Unit 3

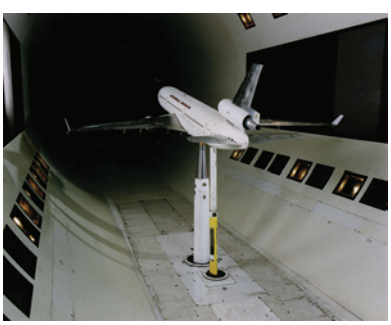
Models for discrete data

Introduction

In Activity 7 of Unit 2, we first looked at the results of an experiment in which the numbers of yeast cells found in each of 400 small squares on a microscope slide were recorded. The data are reproduced in Table 1: the first row of the table gives the number x of yeast cells observed in a square, and the second row of the table gives the number of squares containing x cells for each value of x. The number of yeast cells per square is a random variable, X, say. We decided in Activity 7(b) of Unit 2 that it would be appropriate to model the variation in X using a discrete probability model because each observation is necessarily an integer. Such models are probability distributions, which are just functions describing how the probability of each outcome depends on the value of that outcome.

Table 1 Yeast cells on a microscope slide

Cells in a square, x	0	1	2	3	4	5
Frequency	213	128	37	18	3	1



Model aeroplanes help develop real aeroplanes

The question that was not addressed in Unit 2 is: 'what (discrete) probability model is appropriate for these data?', when we think of the data as a sample from an underlying population. This is a fundamental question that is asked time and time again in statistics and which rarely has an unequivocal answer! We can, however, make considerable progress by considering the nature of the data and how they arose, and by making simplifying modelling assumptions. Remember, however, that we are *modelling* the real world, not seeking exact reproductions of it. The models we use are necessarily simplified versions of reality, but they can be used to great effect to gain insight into situations and to make accurate predictions of what will happen in the future. There's a saying in statistics, attributed to eminent statistician George Box, that 'All models are wrong, but some are useful'.

What can we say about the yeast data? Well, each observation is a count of the number of yeast cells seen in a square. These counts are generally small: many squares have no yeast cells in them at all, quite a lot (but not so many) have one yeast cell in them, and no square has more than five yeast cells in it. The pattern of the frequencies of cells in a square is clearly decreasing as the number of cells increases. So we would expect our probability model to respect this by giving probabilities that also decrease as the number of cells per square increases.

What about the range of X, and hence of the probability model for X? No cell was observed to have six or more cells in it, so does this mean that we can assume that no square can ever have six or more cells in it? If so, we could limit our search for a useful model to those that have range $\{0, 1, 2, 3, 4, 5\}$. Such a model could be useful for some purposes even if the assumption about the end of the range is wrong. But we would probably prefer to allow the model for X to have a range that does allow the possibility of 6 yeast cells in a square, or 7, or 8; where do we stop? Well,

the usual approach would be to say that X has range $\{0, 1, 2, 3, 4, 5, \ldots\}$ (or just $\{0, 1, 2, \ldots\}$), indicating that any (non-negative integer) number of yeast cells is possible in a square. In conjunction with this, we'd make the probabilities associated with such occurrences in our model smaller and smaller as the number of cells increases: for example, 10 yeast cells in a square would be possible under the model but extremely unlikely.

In Unit 2, we looked at probability distributions in general, but did not consider any specific ones.

To make further progress, we need to explore some of the ways in which specific basic probability models (distributions) for discrete data arise. This is the central topic of this unit. In particular, we introduce five specific probability models for discrete data. Three of these relate directly to Bernoulli trials, which are experiments resulting in a binary outcome. These three distributions are the Bernoulli distribution, the binomial distribution and the geometric distribution. We also introduce an important probability model for data in the form of counts, called the Poisson distribution. (It has a less direct connection to Bernoulli trials.) The final discrete distribution that is introduced is one for situations where no value in the range of possible values is more likely than any other value to occur; this is the discrete uniform distribution. It then proves natural, despite the title of the unit, to introduce the analogous distribution for continuous data, the continuous uniform distribution.

The probability distributions mentioned above all have long histories, just a little of which will be touched on in this unit. However, being fundamental to the subject of statistics, they remain as applicable today to modern problems and questions as they were to the contexts in which they were originally conceived, as you will see.

1 Bernoulli trials

The probability models developed in this section and the next two arise in situations where an experiment has just two outcomes: either some event occurs or it does not. For instance: a quality inspector sampling items from a production line will be concerned with whether or not a sampled item is defective; an archer in whether or not her next arrow hits the centre of the target (here, the centre of the target can be defined in terms of the most central one or more rings of the target); a tennis player in whether or not he wins his next match; a medical researcher in whether or not a new drug cures the next patient.

The distribution is named after Swiss mathematician Jakob Bernoulli (1654–1705)

The term **Bernoulli trial** is used to describe a single statistical experiment for which there are two possible outcomes. So taking an item from the production line to determine whether or not it is defective is a Bernoulli trial; and each shot by an archer at a target is a Bernoulli trial – the arrow either hits or misses the centre of the target; and each tennis match may be regarded as a Bernoulli trial in which the tennis player either wins or loses; and so on.

A single Bernoulli trial is so simple a situation that there is only one possible set of probability distributions that can describe it. This is called the **Bernoulli probability distribution**.

1.1 The Bernoulli probability distribution

Where an experiment involves a Bernoulli trial, it is usual to match the number 1 to one outcome and the number 0 to the other; then the outcome of a Bernoulli trial is a random variable with range $\{0, 1\}$. We did this, for example, in Example 7 of Unit 2, where a patient's outcome from a medical treatment was recorded as 1 for 'not cured' and 0 for 'cured'. Similarly, the outcome of tossing a coin might be recorded as 1 for 'heads' and 0 for 'tails', or vice versa. Random variables that can take only two possible values are called *Bernoulli random variables*.

These outcomes are binary.

Suppose that X is the outcome of a single Bernoulli trial, taking the value 0 or 1. Then X has a Bernoulli probability distribution, or **Bernoulli distribution** for short. If we let p denote the probability that X takes the value 1, then X is said to have a Bernoulli distribution *with parameter* p. Since

$$P(X = 1) + P(X = 0) = 1,$$

the probability mass function of X is

$$P(X = 1) = p(1) = p \quad \text{and} \quad P(X = 0) = p(0) = 1 - p.$$

Here, as usual, $p()$ denotes the p.m.f., but the letter p is also used conventionally for the Bernoulli parameter.

That is,

$$p(x) = \begin{cases} 1 - p & x = 0 \\ p & x = 1. \end{cases}$$

Because a probability mass function has the property $p(x) > 0$ for each x in the range of X, it must be the case that $0 < p < 1$.

All Bernoulli distributions have this form so there is a *family* of Bernoulli distributions – the value of p determines a particular member of the family. The parameter p is thereby said to *index* the family of Bernoulli distributions. Statisticians are not especially consistent in their terminology here: either an individual member of the Bernoulli family of distributions (with a particular value chosen for p) or the whole family of Bernoulli distributions can be referred to as 'the Bernoulli distribution'.

The p.m.f. of a Bernoulli distribution may also be written in multiplicative form as

$$p(x) = p^x(1 - p)^{1-x}, \quad x = 0, 1.$$

As will become apparent in Subsection 2.1, this gives $p(0)$ and $p(1)$ in a single expression that can be related to the next distribution that we consider.

Activity 1 *Does the formula work?*

Check that the equation above gives $p(1) = p$ and $p(0) = 1 - p$.

The Bernoulli distribution is summarised in the following box.

The Bernoulli probability model

A discrete random variable X with range $\{0, 1\}$ is said to have a **Bernoulli distribution with parameter p**, where $0 < p < 1$, if it has probability mass function

$$p(x) = \begin{cases} 1-p & x = 0 \\ p & x = 1 \end{cases}$$

or equivalently

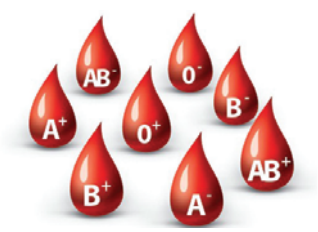
$$p(x) = p^x(1-p)^{1-x}, \quad x = 0, 1. \tag{1}$$

The symbol '$\sim$' is read 'is distributed as' or simply 'is'.

This is written $X \sim \text{Bernoulli}(p)$.

The model may be applied to the outcome of a Bernoulli trial where the probability of obtaining the outcome 1 is equal to p.

Example 1 *Blood groups*



In a sample of people living in London who suffered from peptic ulcers, 911 had blood group O and 579 had blood group A. (Source: Woolf, B. (1955) 'On estimating the relation between blood group and disease', *Annals of Human Genetics*, vol. 19, no. 4, pp. 251–3.) If one of these people is picked at random, then the probability that this person has blood group O is $911/(911 + 579) = 911/1490 \simeq 0.611$. So, within this sample, the random variable which takes the value 1 if a person is group O and 0 if a person is group A has a Bernoulli distribution with parameter $p = 0.611$. (For simplicity, we assume that p is exactly 0.611.) The p.m.f. of this distribution is

$$p(x) = \begin{cases} 0.389 & x = 0 \\ 0.611 & x = 1 \end{cases}$$

or equivalently

$$p(x) = (0.611)^x\,(0.389)^{1-x}, \quad x = 0, 1.$$

Activity 2 *Norwegian women*

In a study of all 6503 women aged between 35 and 49 in the Norwegian county of Sogn og Fjordane, 591 of the women were found to have passed the menopause. (Source: Keiding, N. (1991) 'Age-specific incidence and prevalence: a statistical perspective', *Journal of the Royal Statistical Society, Series A*, vol. 154, no. 3, pp. 371–412.)

(a) If one of these women is chosen at random, calculate the probability that she has passed the menopause.

(b) Define a suitable random variable X to indicate whether a woman randomly chosen from this population has passed the menopause, and write down its probability mass function.

Exercise on Section 1

Exercise 1 *Faulty street lamps*

In Activity 8(a) of Unit 2, you defined the random variable X to be 1 if a randomly chosen street lamp from a consignment of street lights is faulty and to be 0 if a randomly chosen street lamp from the consignment is not faulty. You wrote its probability mass function as

$$p(x) = \begin{cases} 0.96 & x = 0 \\ 0.04 & x = 1. \end{cases}$$

Give a name and parameter value to the distribution of which this is the p.m.f., and write the p.m.f. in multiplicative form.

2 The binomial probability distribution

In Section 1, the Bernoulli distribution was introduced as the probability model for the outcome of a single Bernoulli trial. Situations are often encountered, however, in which there is not just a single Bernoulli trial, but a set of such trials. For example, in evaluating a new drug for a particular condition we would not wish to base conclusions on the reaction of a single patient. Instead, we would treat many patients with the condition and look at the proportion for whom the treatment was beneficial. We would use this proportion as an estimate of the probability, p say, that the drug would be beneficial for a randomly selected patient with that condition.

Note, however, that there are two implicit assumptions in this.

- We have assumed that whether or not one patient responds favourably to the drug does not affect the probability that another will respond favourably. In general, if it is valid to assume that the outcome of one trial does not influence the probability of the outcome of another trial, then the trials are said to be **independent**. So, using this terminology, the assumption above is that the responses of the patients to the drug are independent.
- We have also assumed that the probability that one patient with the condition responds favourably to the drug is the same as the probability that another patient with the condition will respond favourably to the drug. That is, p is the same for all the Bernoulli trials in the set of such trials. This may not be so: effectiveness of the treatment may depend on other factors such as the patient's age or severity of illness. It is, however, often reasonable to assume that p is the same for all patients in a group defined by, for example, patients of similar age and severity of illness.

So the treatment of each patient with the condition (and in a well-defined group of similar patients) is modelled as a Bernoulli trial with the same parameter p; and these trials are presumed to be independent of one another. The main quantities of interest are the total number of patients

who respond favourably in a sample of patients treated with the drug (the total number of successful trials) and the proportion of patients who respond favourably (the proportion of successful trials, which is the total number of successful trials divided by the total number of patients). This is illustrated in the next example.

Example 2 *Headache relief*

In a sample of eight patients, five responded successfully to a treatment to relieve headache, while the other three failed to respond. The total number of patients in the sample who responded successfully to the treatment is therefore 5. The proportion of patients in the sample who responded successfully to the treatment is 5/8. This number provides an estimate of the proportion of successful responses in the population of headache sufferers given the treatment.

The situation where the random variable of interest is the *total number of successful trials* in a set of independent Bernoulli trials is a very common one in statistics. This number is a random variable, X say. The probability model for this discrete random variable is one of the standard models that we will be using on many occasions in this module. Its definition is as follows.

The binomial distribution

If the probability of success in each of a set of n independent Bernoulli trials has the same value p, then the random variable X, which represents the total number of successes in the n trials, is said to have a **binomial distribution with parameters n and p**. This is written $X \sim B(n, p)$.

The expressions 'X is a binomial random variable', 'X is binomially distributed' and 'X is binomial' are sometimes used. As with the Bernoulli distribution, all binomial distributions have the same form and there is a family of binomial distributions, the values of n and p determining a particular member of the family. Again, statisticians use the terminology 'the binomial distribution' both to refer to an individual member of the binomial family of distributions (with particular values chosen for n and p) and for the whole family of binomial distributions (indexed by n and p).

The same lack of distinction will apply to other distributions considered later in the module.

Before finding the probability mass function of the binomial probability distribution, let us look briefly at some situations where it might provide an adequate model and some where it would not.

Example 3 *Coin tossing*

Suppose that a fair coin is tossed three times. Each time it may come up 'heads' or 'tails', and each of these outcomes has probability 1/2. It is reasonable to assume that the trials are independent: that is, that the

Here, a head is treated as a 'success', a tail as a 'failure'.

outcome of the first toss (heads or tails) does not influence the chance of the coin coming up heads or tails on the next toss, or on any subsequent toss. So these three trials satisfy the conditions for the total number of heads in the three tosses to have a binomial distribution with parameters $n = 3$ and $p = 1/2$. That is, the probability distribution for the total number of heads is binomial, $B(3, 1/2)$.

Example 4 *Political attitudes in one household*

In a survey of political attitudes in a two-party system, we might ask all the adults in a household how they intend to vote. There are two possible outcomes. For each adult in the household, we can define a random variable X, which takes the value 1 if that person says they intend to vote RED and the value 0 if they say they intend to vote BLUE. For any one adult, this random variable has a Bernoulli distribution. Then we could add up these random variables to produce a new one, Y, which is the total number of adults in the household who say they intend to vote RED. However, adults within a household are likely to influence each other in the way they vote, so the X scores would not be independent thus the random variable Y would *not* have a binomial distribution.

Activity 3 *Might a binomial distribution be a good model?*

In each of the following scenarios, consider whether or not a binomial distribution might be a good model for the total number of individuals or items in the sample with the attribute in question. Give reasons for your answers.

(a) In a study of ethnicity, 50 households containing two people were randomly sampled from those in a large UK city, and whether or not the 100 selected people were of Indian descent was recorded.

(b) Twenty children were randomly sampled from schools in a number of similar inner-city estates and whether or not they were obese was recorded.

(c) In a study of sporting activity, all young people living in a particular street in the city were asked whether or not they play football.

By the way, 'successes' and 'failures' are the standard ways of referring to the 1s and 0s that are the outcomes of a collection of Bernoulli trials. This is so whether or not a success is really a success! For instance, in several examples in Unit 2, we considered a Bernoulli trial in which '1' corresponded to a person who was not cured by a medical treatment, '0' to someone who was cured; in Activity 3(b), a 'success' would be 'being obese'. The success/failure terminology is just a standard convenience, 'success' being a name for the focus of what is being studied; you should always report outcomes and interpret results in language that is meaningful in the context at hand.

"I *am* dressed for success! Of course, my idea of success may not be exactly the same as yours."

2.1 The binomial probability mass function

It's a p.m.f., not a p.d.f., because the binomial is a discrete distribution.

In this subsection, we will derive the probability mass function for the binomial distribution, $B(n, p)$. Let us start by considering its range, that is, all the possible outcomes of a binomial random variable.

Example 5 *Tossing a coin three times*

In Example 3, we considered tossing a coin three times and counting the number of heads. It was decided that the number of heads could be modelled by a binomial distribution with parameters $n = 3$ and $p = 1/2$. The possible numbers of heads arising in three tosses of a coin are 0, 1, 2 or 3. So the range of the $B(3, 1/2)$ distribution is $\{0, 1, 2, 3\}$.

Activity 4 *The range of the total score*

(a) An experiment is performed consisting of a sequence of 15 independent Bernoulli trials. If a trial is successful, a score of 1 is recorded; otherwise, a score of 0 is recorded. The probability that a trial is successful is p. The total score for the experiment, X, is obtained by adding together the scores recorded for all 15 trials. We therefore know that $X \sim B(15, p)$. What is the range of the $B(15, p)$ distribution?

(b) Generalising part (a), suppose the sequence of independent Bernoulli trials is of length n. The probability that a trial is successful is still p, and the total score for the experiment, X, is distributed as $B(n, p)$. What is the range of the $B(n, p)$ distribution?

So from Activity 4(b), we have the range of the $B(n, p)$ distribution. But how can we calculate probabilities for a binomial model? To illustrate the calculations, consider the following example.

Example 6 *Leaving London*

In recent years, there has been a considerable increase in working people leaving London for new jobs elsewhere; this trend is at least partly driven by financial considerations, especially the high cost of housing in London. Suppose, therefore, that we are interested in how satisfied working age people are with living in London. To investigate this, a random sample of people of working age living in London could be drawn, and each person in the sample asked questions such as whether or not they are actively considering taking a job elsewhere. Asking a randomly selected person whether or not they are seeking a job outside London is a Bernoulli trial. We will assume that whether or not one person in the sample is considering leaving London does not affect the probability that any other person in the sample is considering leaving London, that is, we will assume that the trials are independent. We will also assume that the probability that a person of working age is considering leaving London is the same for everyone. Suppose also that 1 in 3 people will answer Yes to the question, so that the probability that a randomly chosen person will answer Yes is $\frac{1}{3}$. (In reality, estimating this probability is the purpose, or one of the purposes, of such an investigation.)

'An Ipsos Mori poll for London Councils found more than a third of all people living in the capital were actively considering taking a job elsewhere because of high housing costs' (*Evening Standard*, 22 October 2015)

Suppose that three people are asked the question: 'Are you actively considering taking a job outside London?' The number of people who answer Yes has a binomial distribution, $B(3, \frac{1}{3})$. What is the probability that two out of the three people will answer Yes?

There are two stages involved in the calculation of this probability.

First, consider the outcome where the first person answers Yes, the second answers Yes, and the third person answers No. The probability that the first person answers Yes is $\frac{1}{3}$. Since we are assuming that p is the same for everyone, the probability that the second person answers Yes is also $\frac{1}{3}$, and the probability that the third answers No is $\frac{2}{3}$. Also, since we are assuming that whether or not a person in the sample is considering leaving London is independent of whether or not any other person in the sample is considering leaving London, the overall probability that the responses are

Yes Yes No,

in that order, is given by

$$P(\text{Yes Yes No}) = P(\text{Yes}) \times P(\text{Yes}) \times P(\text{No}) = \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{2}{3} = \tfrac{2}{27}.$$

Here we are using the probability rule for multiple independent events from Subsection 1.1 of Unit 2.

Now we move to the second stage of the calculation. In pursuing this survey, we are actually interested not in recording the order in which the responses occurred, only in counting the number of responses of each type. We require the probability that, of the three people questioned, two say Yes and one says No. There are exactly three different ways this could happen, and their probabilities are as follows:

$$P(\text{Yes Yes No}) = P(\text{Yes}) \times P(\text{Yes}) \times P(\text{No}) = \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{2}{3} = \tfrac{2}{27},$$

$$P(\text{Yes No Yes}) = P(\text{Yes}) \times P(\text{No}) \times P(\text{Yes}) = \tfrac{1}{3} \times \tfrac{2}{3} \times \tfrac{1}{3} = \tfrac{2}{27},$$

$$P(\text{No Yes Yes}) = P(\text{No}) \times P(\text{Yes}) \times P(\text{Yes}) = \tfrac{2}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} = \tfrac{2}{27}.$$

If we now disregard the order of the responses, we find that the probability of receiving two Yes responses and one No response is

$$P(\text{Yes Yes No}) + P(\text{Yes No Yes}) + P(\text{No Yes Yes})$$
$$= \tfrac{2}{27} + \tfrac{2}{27} + \tfrac{2}{27} = 3 \times \tfrac{2}{27} = \tfrac{6}{27} = \tfrac{2}{9}.$$

The pluses arise from the elementary probability result that for mutually exclusive events E_1, E_2 and E_3, $P(E_1 \text{ or } E_2 \text{ or } E_3) = P(E_1) + P(E_2) + P(E_3)$.

Notice that the first stage in calculating the probability in this example was to find the probability of two Yeses and one No in that order: this probability was $\frac{2}{27}$. You then saw that the probability of two Yeses and one No in each of any other order was also $\frac{2}{27}$. So to find the required probability it was necessary to count the number of different ways of ordering two Yeses and one No: three, in this case. Hence the required probability was $3 \times \frac{2}{27}$.

We can use this method to find other similar probabilities. For instance, what is the probability that in a sample of ten people questioned, seven respond Yes?

First, we find the probability of 7 Yeses and 3 Noes in that order: this is

$$\tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{2}{3} \times \tfrac{2}{3} \times \tfrac{2}{3} = \left(\tfrac{1}{3}\right)^7 \times \left(\tfrac{2}{3}\right)^3.$$

Notice, however, that the probability of any particular arrangement, or *combination*, of 7 Yeses and 3 Noes is also equal to $\left(\frac{1}{3}\right)^7 \times \left(\frac{2}{3}\right)^3$; for instance,

Such a combination is sometimes called an unordered sequence

$$P(\text{Yes Yes Yes No No Yes Yes Yes No Yes})$$
$$= \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{2}{3} \times \tfrac{2}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{2}{3} \times \tfrac{1}{3} = \left(\tfrac{1}{3}\right)^7 \times \left(\tfrac{2}{3}\right)^3.$$

So, to complete the calculation, we must find the number of different ways of ordering 7 Yeses and 3 Noes. If you try to list all of them, you will soon realise that there is a large number of combinations of 7 Yeses and 3 Noes. However, it is not necessary to list all the possible sequences of responses: there is a general formula which can be used to calculate the number of such sequences. This formula is stated in the following box.

Number of combinations of objects of two types

The number of different ways of ordering x objects of one type and $n - x$ objects of a second type in a sequence of n objects is given by

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}. \qquad (2)$$

This formula holds for any integer value of x from 0 to n. The number $\binom{n}{x}$ is read 'n choose x'.

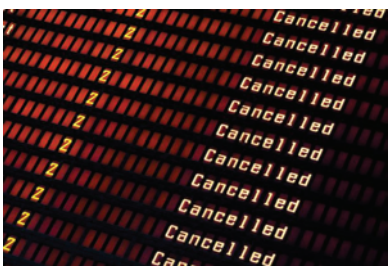
Alternative notation for $\binom{n}{x}$ is ${}^nC_x$, sometimes read as 'n c x'

The number $x!$ is read 'x factorial'. For any positive integer x, the notation $x!$ is shorthand for the number $1 \times 2 \times 3 \times \cdots \times x$. The number 0! is defined to be 1.

To illustrate the use of the formula, we will find the number of different ways of obtaining 7 Yeses (and 3 Noes) from a sample of ten people questioned. In this case, $n = 10$ and $x = 7$ in Equation (2) (and hence $n - x = 3$), so the number required is

$$\binom{10}{7} = \frac{10!}{7!\,3!} = \frac{1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8 \times 9 \times 10}{(1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7)(1 \times 2 \times 3)}$$
$$= \frac{8 \times 9 \times 10}{1 \times 2 \times 3} = 120.$$

The quantities $\binom{n}{x}$ are often called *binomial coefficients*. If you have not previously calculated values of binomial coefficients, then you may find it helpful to work through the next activity. Happily, as above, evaluating binomial coefficients is always simplified by cancellation of terms in the numerator and the denominator.



Lots more cancellation

Activity 5 *Binomial coefficients*

Use Equation (2) to find the values of the following binomial coefficients.

(a) $\binom{5}{3}$ (b) $\binom{7}{1}$ (c) $\binom{8}{0}$ (d) $\binom{6}{4}$

Now we can complete the calculation of the probability of obtaining 7 Yes responses and 3 No responses from a sample of ten people when the probability of a Yes response is $\frac{1}{3}$. It is

$$\binom{10}{7} \times \left(\tfrac{1}{3}\right)^7 \times \left(\tfrac{2}{3}\right)^3 = 120 \times \left(\tfrac{1}{3}\right)^7 \times \left(\tfrac{2}{3}\right)^3 \simeq 0.016.$$

This is the probability that a binomial random variable with parameters $n = 10$ (the sample size) and $p = \frac{1}{3}$ (the probability of obtaining a Yes response) will take the value 7. That is, if $X \sim B\left(10, \frac{1}{3}\right)$, then

$$P(X = 7) = \binom{10}{7} \left(\tfrac{1}{3}\right)^7 \left(\tfrac{2}{3}\right)^3.$$

This is just a special case of the following result: if $X \sim B(10, \frac{1}{3})$, then

$$P(X = x) = \binom{10}{x} \left(\tfrac{1}{3}\right)^x \left(\tfrac{2}{3}\right)^{10-x}, \quad x = 0, 1, 2, \ldots, 10.$$

This formula arises because there are $\binom{10}{x}$ combinations of x Yeses out of 10 Yeses and Noes, and the probabilities of x Yeses and $10 - x$ Noes are $\left(\frac{1}{3}\right)^x$ and $\left(\frac{2}{3}\right)^{10-x}$, respectively. So, for example, taking x equal to 0 in this formula gives

$$P(X = 0) = \binom{10}{0} \left(\tfrac{1}{3}\right)^0 \left(\tfrac{2}{3}\right)^{10} = \frac{10!}{0!\,10!} \left(\tfrac{2}{3}\right)^{10} = \left(\tfrac{2}{3}\right)^{10} \simeq 0.017.$$

This is the probability that all ten responses are No.

Of course, the results just obtained do not apply only to sequences of Yeses and Noes. The method used to find the probability of obtaining 7 Yeses (and 3 Noes) in a sample of 10 Yes–No responses can be generalised to find the probability mass function for a binomial distribution with parameters n and p, because this gives the probability of x successes (1s) (and $n - x$ failures, 0s) in a sample of n independent Bernoulli trials each with probability of success p. In this general situation, $\binom{n}{x}$ is the number of different ways of obtaining x successes and $n - x$ failures in a sequence of n Bernoulli trials. Also, the probabilities of x successes and $n - x$ failures are p^x and $(1 - p)^{n-x}$, respectively. By multiplying together $\binom{n}{x}$, p^x and $(1 - p)^{n-x}$ we have the probability mass function of the binomial distribution, as given in the box below.

The binomial probability model

If a random variable X has a binomial distribution with parameters n and p, where $0 < p < 1$, then it has probability mass function

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n. \tag{3}$$

This is written $X \sim B(n, p)$.

The binomial distribution provides a probability model for the total number of successes in a sequence of n independent Bernoulli trials, in which the probability of success in a single trial is p.

Example 7 *Calculating binomial probabilities*

Suppose that $Y \sim B(5, 0.6)$. Equation (3) may be used with $n = 5$ and $p = 0.6$ to find probabilities involving Y. For example,

$$P(Y = 3) = \binom{5}{3} (0.6)^3 (1 - 0.6)^{5-3} = \frac{5!}{3!\,2!} (0.6)^3 (0.4)^2$$
$$= 10(0.6)^3 (0.4)^2 = 0.3456.$$

Activity 6 *More binomial probabilities*

(a) If $X \sim B(6, 0.4)$, find the probability $P(X = 4)$.

(b) If $V \sim B(8, 0.3)$, find the probability $P(V = 2)$.

Activity 7 *Patient dropout*

Suppose that a study is undertaken to compare the safety and efficacy of two antidepressant drugs. Eighteen patients are each randomly allocated to one of three groups; there are six to a group. The first group is treated with Drug A and the second with Drug B. Patients in the third group are treated with a placebo.

A placebo is a substance which contains no active medication, but which is given to the patients in the same way as the treatments being studied, so that the analysis can be controlled for any natural remission, called the placebo effect.

One of the problems associated with studies of this sort is that patients occasionally drop out: they cease treatment before the study is completed. This might be for reasons unrelated to their course of treatment, or because they suffer from side effects, or it might be because they perceive no beneficial effect from their treatment. Consequently, the phenomenon is a complicating feature in a statistical analysis of the results of such studies.

A previous study suggests that the percentage of patients in placebo groups who drop out might be about 14%. (Source: Dunbar, G.C. et al. (1991) 'A comparison of paroxetine, imipramine and placebo in depressed outpatients', *British Journal of Psychiatry*, vol. 159, pp. 394–8.)

(a) Using this estimate for the value of the parameter p in a binomial model, calculate the probabilities of the following events for the placebo group in the present study.

(i) All six patients drop out.

(ii) None of the six drops out.

(iii) Exactly two from the group drop out.

(b) An assumption of the binomial model is that of independence from trial to trial. Interpret this assumption in the context of the study, and comment on whether you believe that, in this case, it is a reasonable assumption.

Three typical binomial probability mass functions are shown in Figure 1. The three distributions illustrated are: (a) $B\left(3, \frac{1}{2}\right)$, (b) $B\left(10, \frac{3}{4}\right)$ and (c) $B(6, 0.14)$. Figure 1(a) shows the probability mass function for the number of heads obtained when a fair coin is tossed three times (see Example 3). Figure 1(b) shows the probability mass function for the number of arrows that hit the centre of a target when, for example, an archer shoots ten arrows and the probability that each arrow she shoots hits the centre of the target is 3/4. Figure 1(c) shows the probability mass function for the number of patients (out of a group of six) who drop out of a study of the efficacy of antidepressant drugs (see Activity 7).

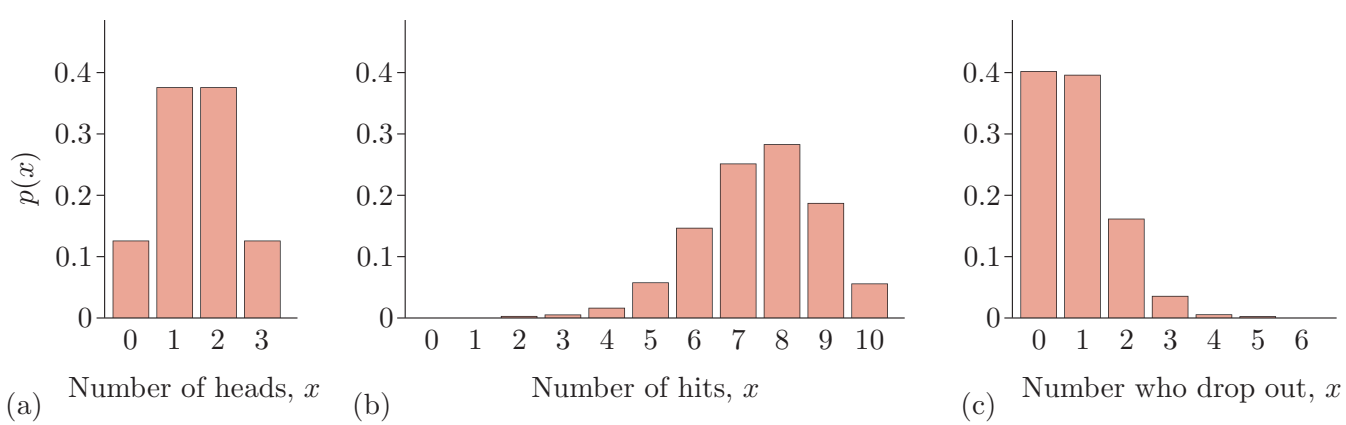


Figure 1 Typical members of the binomial family: (a) $B\left(3, \frac{1}{2}\right)$, (b) $B\left(10, \frac{3}{4}\right)$, (c) $B(6, 0.14)$

Activity 8 *The shape of binomial distributions*

Look at the diagrams in Figure 1. Comment on the shapes of these binomial probability mass functions.

In fact, the comments in the solution to Activity 8 apply to all binomial distributions. The shapes of binomial probability mass functions are explored in more detail in the following screencast.

Screencast 3.1 The shapes of binomial distributions

The Bernoulli distribution is a special case of the binomial distribution.

Activity 9 *Bernoulli as binomial*

(a) By considering Bernoulli trials, to what binomial distribution do you think the Bernoulli distribution with success probability p is equivalent?

(b) Confirm the result of part (a) by showing that Equation (3) for the p.m.f. of the binomial distribution reduces to Equation (1) for the p.m.f. of the Bernoulli distribution.

2.2 More binomial probabilities

In Subsection 2.1, we considered the calculation of probabilities that a binomially distributed random variable takes a particular value. In this subsection, we consider the calculation of probabilities that a binomially distributed random variable takes one of a set of values. Since we are dealing with a discrete distribution, these probabilities can be obtained by addition.

Example 8 *Patient dropout*

In Activity 7, it was argued that the number of patients who drop out of a group of 6 patients receiving a placebo in a drug trial, Y, might be modelled as $Y \sim B(6, 0.14)$. In Activity 7(a)(iii), you calculated that the probability that exactly two people drop out is $P(Y = 2) \simeq 0.161$. What is the probability that more than two people drop out?

We need $P(Y > 2)$, and the most direct approach would be to calculate

$$P(Y > 2) = P(Y = 3) + P(Y = 4) + P(Y = 5) + P(Y = 6).$$

This involves calculating and adding four individual binomial probabilities. We can make the calculations easier by instead using the fact that

$$P(Y > 2) = 1 - P(Y \leq 1) = 1 - \{P(Y = 0) + P(Y = 1)\};$$

this involves calculating and adding just two individual binomial probabilities. We have

$$\begin{aligned} P(Y > 2) &= 1 - \{P(Y = 0) + P(Y = 1)\} \\ &= 1 - \left\{ \binom{6}{0} (0.14)^0 (1 - 0.14)^6 + \binom{6}{1} (0.14)^1 (1 - 0.14)^5 \right\} \\ &= 1 - \left\{ \frac{6!}{0!\,6!} (0.86)^6 + \frac{6!}{1!\,5!} (0.14)(0.86)^5 \right\} \\ &= 1 - \left\{ (0.86)^6 + 6(0.14)(0.86)^5 \right\} \\ &\simeq 1 - (0.4046 + 0.3952) \simeq 0.200. \end{aligned}$$

(As usual, it is good practice to retain more decimal places in intermediate calculations than in the final result.)

Activity 10 *Calculating binomial probabilities*

(a) If $W \sim B(7, 0.6)$, find the probability $P(W \leq 1)$.

(b) In Example 6, we considered the number of people, Y say, who were considering leaving London, out of a random sample of size 3. The distribution of Y was modelled as $B\left(3, \frac{1}{3}\right)$. What is the probability that either none of, or all of, the people in the sample were considering leaving London?

(c) If $X \sim B(6, 0.8)$, find the probability $P(X > 4)$.

Despite the name, this IT firm will not be able to help!

Example 8 and Activity 10 concerned examples where we needed to calculate only a few individual binomial probabilities. This was feasible by hand calculation. Some binomial calculations require rather more individual binomial probabilities, however, as in the next example.

Example 9 *Multiple choice examination scores*

One of the components of assessment for the statistics students at a particular British university is a multiple choice examination consisting of 20 questions. For each question the correct answer is one of five options. Students indicate which one of the five options they believe to be correct. Sometimes some of the students give the impression that they have gone through the paper guessing answers at random. If 1 mark is awarded for each correct answer and the pass mark is 10, what is the probability that a student who guesses answers at random will pass the examination?

Since such a student guesses answers at random, an answer to any particular question is independent of an answer to any other question. Moreover, since there are five possible answers to each question and since the selection is made at random, the probability of picking the correct option is 1/5 for each question. Thus the answers of a student who guesses at random form a sequence of 20 independent Bernoulli trials, each with probability of success 1/5 or 0.2. So the total number of correct answers

given to the 20 questions is a random variable having a binomial distribution with parameters $n = 20$ and $p = 0.2$. That is, if the random variable T denotes the total number of correct answers, then $T \sim B(20, 0.2)$.

If the pass mark is 10, then the probability that a student who guesses answers will score less than 10 (and so fail the examination) is given by

$$P(T < 10) = P(T \leq 9) = F(9),$$

where F is the cumulative distribution function of T. Recall from Subsections 4.1 and 4.2 of Unit 2 that the c.d.f. of any discrete distribution is calculated by addition, so

$$P(T \leq 9) = P(T = 0) + P(T = 1) + P(T = 2) + \cdots + P(T = 9).$$

This requires calculating and adding ten binomial probabilities. These calculations (and more) have been done and are presented in Table 2, which gives the probability mass function $p(t) = P(T = t)$ and the cumulative distribution function $F(t) = P(T \leq t)$ for the binomial random variable $T \sim B(20, 0.2)$. (All the probabilities except $F(20)$, which is exact, have been rounded to four decimal places.)

Table 2 The probability distribution of $T \sim B(20, 0.2)$

t	$p(t)$	$F(t)$
0	0.0115	0.0115
1	0.0576	0.0692
2	0.1369	0.2061
3	0.2054	0.4115
4	0.2182	0.6297
5	0.1746	0.8042
6	0.1091	0.9133
7	0.0546	0.9679
8	0.0222	0.9900
9	0.0074	0.9974
10	0.0020	0.9994
11	0.0005	0.9999
12	0.0001	1.0000
13	0.0000	1.0000
14	0.0000	1.0000
15	0.0000	1.0000
16	0.0000	1.0000
17	0.0000	1.0000
18	0.0000	1.0000
19	0.0000	1.0000
20	0.0000	1

So if the pass mark is 10, then the probability that a student who randomly guesses his answers will fail is

$$P(T \leq 9) = F(9) = 0.9974,$$

and the probability that he will pass is

$$P(T \geq 10) = 1 - P(T \leq 9) = 1 - F(9) = 1 - 0.9974 = 0.0026.$$

At this juncture, you might expect to be provided with a formula for the c.d.f. of the binomial distribution. Unfortunately, the c.d.f. of a binomial random variable does not have a convenient mathematical form. The options for calculating the binomial c.d.f. are therefore calculation by 'hand' (i.e. calculator) by adding individual binomial probabilities, or using your computer to make the calculations for you. So you will next use Minitab to calculate probabilities for binomial distributions. The activities associated with this endeavour will give you further practice at deciding what probability you need to find to solve a problem.

Refer to Chapter 7 of Computer Book A for the work on calculating binomial probabilities using Minitab.

Although the c.d.f., $F(x)$, of the $B(n, p)$ distribution does not have a nice mathematical form, it *is* possible to check that $\sum p(x) = 1$ and hence – at this very late stage! – that Equation (3) defines a valid p.m.f. The range of the binomial distribution is $\{0, 1, 2, \ldots, n\}$ so the summation in question is $\sum_{x=0}^{n} p(x)$. The key to showing that $\sum_{x=0}^{n} p(x) = 1$, as you will see in the next activity, is the *binomial theorem* of mathematics, which can be written

Also, individual binomial probabilities are non-negative.

$$(a + b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}.$$

It is this link with the binomial theorem that gives the binomial distribution its name, and why the terms $\binom{n}{x}$ are called binomial coefficients.

Activity 11 *Proving that $\sum p(x) = 1$ for the binomial distribution*

For the $B(n, p)$ distribution,

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

Use the binomial theorem to confirm that $\sum_{x=0}^{n} p(x) = 1$.

You should by now be familiar with the fact that the binomial distribution is indexed by two parameters, n, the number of independent Bernoulli trials, and p, the probability of success in each trial. You might also have realised that the two parameters are not always on an equal footing, in the following sense. In every situation explored in this section, the parameter n is a known quantity; in each case, we know how many trials there are in the study or experiment. The value of p, on the other hand, is often unknown and would need to be estimated in practice. Examples include the probability that a working-age person is considering leaving London because of high housing costs, which we assumed to be 1/3 in Example 6, and the probability that a patient on a placebo in a drug trial drops out of the trial, which we took to be 0.14 on the basis of a previous study in Activity 7.

Exceptions include $p = 1/2$ for an unbiased coin and $p = 1/g$ for guessing from among g options.

Exercises on Section 2

Exercise 2 *More binomial coefficients*

Find the values of the following binomial coefficients.

(a) $\binom{6}{1}$ (b) $\binom{9}{2}$ (c) $\binom{7}{4}$

Exercise 3 *Binomial probabilities*

(a) If $X \sim B(4, 0.6)$, find the probability $P(X = 3)$.

(b) If $Y \sim B(7, 0.2)$, find the probability $P(Y = 2)$.

(c) If $Z \sim B\left(8, \frac{3}{4}\right)$, find the probability $P(Z > 6)$.

Exercise 4 *More binomial probabilities*

(a) The probability that an item from a production line is defective is 0.05. Find the probability that a sample of 20 items will contain at most one defective item.

(b) The probability that an archer hits the centre of the target with each arrow that she shoots is 0.9. Find the probability that eight out of ten arrows that she shoots hit the centre of the target.

(c) The probability that a tennis player wins each match he plays against a friend is 0.7. If he plays five matches, find the probability that he wins at least three of them.

3 The geometric probability distribution

In Section 1, a Bernoulli trial was defined to be a statistical experiment in which exactly one of two possible outcomes occurs. The outcomes may be, for instance, Success–Failure, Yes–No, On–Off or Male–Female. It is usual to identify one of the outcomes (success) with the number 1 and the other (failure) with the number 0. If p is the probability that a single trial results in a success, then the outcome of a single trial is a random variable which has a Bernoulli distribution with parameter p $(0 < p < 1)$. Also, the total number of successes in a sequence of n independent trials, each with the same probability of success p, is a random variable that has a binomial distribution with parameters n and p, where n is the total number of trials.

In this section another probability model associated with a sequence of independent Bernoulli trials is introduced; this is a model for a random variable which represents the number of trials up to and including the *first* success. This model – the **geometric distribution** – is developed in Subsection 3.1.

Over the last three hundred years, the pattern of boys and girls in families has been studied by many people. Some theories and results of these investigations are discussed briefly in Subsection 3.2. In particular, a model for family patterns based on Bernoulli trials is described. The binomial and particularly the geometric distributions will be used.

3.1 The geometric probability model

In Examples 10 and 11, two situations which may be modelled by a sequence of Bernoulli trials are described, in each of which the same question is of central importance: how many trials do we need to observe until we get a success? Example 10 revisits Example 6 of Unit 2; Example 11 is new.

Example 10 *Waiting to join in*

In many board games, progress round the board is determined by the score obtained when a six-sided die is rolled. In some games, you cannot start playing until you have obtained your first six (and then you move

accordingly). If you score some other number, you have to wait until your next turn in order to make another attempt to obtain a six. You may therefore start playing with your first roll of the die, or your second, or third, and so on. The number of rolls of the die until you start playing is a random variable with range $\{1, 2, 3, \ldots\}$.

Example 11 *Silicon chips*

The manufacture of silicon chips is an extremely sensitive operation, requiring engineering accuracies many orders greater than those required in most other manufacturing contexts, and a working environment that is clinically clean. In the early days of chip technology, most chips were defective – they did not work properly (or often, they did not work at all). Either they were dirty (a speck of dust just 0.5 microns across can cause havoc on a circuit board where the tracks carrying current are only 0.3 microns across), or not all the connections were correctly made. At all stages of slicing, lapping, etching, cleaning and polishing involved in the manufacture of a chip, defective units are identified and removed. Even so, possibly as many as one chip in 20 is faulty.

A factory making silicon chips

During the manufacturing process, there may be 'runs of rough' – intervals during which nothing seems to go very well and the product defective rate is rather high. These periods alternate with intervals where there are fewer defectives than average. However, in what follows we will make the simplifying assumption that chip quality can be regarded as invariant and independent from chip to chip.

Suppose that a quality inspector at a silicon chip factory introduces a new quality test. At random times she samples completed chips from the assembly line. She makes a note of the number of chips sampled up to and including the first defective chip she finds. If this number reaches or exceeds some predetermined tolerance limit, then she assumes that factory procedures are running efficiently. Otherwise the production process is stopped for assessment and readjustment, since it appears that defectives are occurring too frequently. The number of chips sampled is a random variable with range $\{1, 2, 3, \ldots\}$.

In both these examples the same type of random variable is being counted: essentially, the number of trials from the start of a sequence of independent Bernoulli trials up to *and including* the first success. Notice that the trial at which that success occurs is included in the count. In Example 10, success is obtaining a six on the roll of a six-sided die; and in Example 11, success is the identification of a defective chip.

Recall that 'success' is usually the version of the outcome that is the focus of the investigation.

If the number of trials up to and including the first success is denoted by X, then

$X = 1$ if the first trial is a success,

$X = 2$ if the first trial is a failure and the second is a success,

$X = 3$ if the first two trials are failures and the third is a success,

and so on.

Activity 12 *The probability function of X*

The number X is a random variable: at the start of the sequence it is impossible to forecast with certainty the number of the trial at which the first success will occur. You may assume that the trials are independent and that the probability of success in each trial is p, where $0 < p < 1$.

(a) Write down the probability $P(X = 1)$.

(b) Write down the probability $P(X = 2)$.

(c) Write down the probability $P(X = 3)$.

(d) Using your answers to parts (a), (b) and (c), suggest a general formula for the probability $P(X = x)$.

To recap the result of Activity 12(d), if the first success is on the xth trial, then the first $x - 1$ trials must each be a failure and the xth trial must be a success. The probability that each of the first $x - 1$ trials is a failure is

$$\underbrace{(1-p) \times (1-p) \times \cdots \times (1-p)}_{x-1 \text{ terms}} = (1-p)^{x-1}.$$

The probability that the xth trial is a success is p. Hence the probability that the xth trial is the *first* success is $(1-p)^{x-1} \times p$. So if the random variable X is the number of the Bernoulli trial on which the first success occurs, then the distribution of X is called the geometric distribution and we have just obtained its probability mass function. The geometric distribution is defined more formally in the following box.

The geometric distribution

Suppose that in a sequence of independent Bernoulli trials, the probability of success is constant from trial to trial and equal to p, where $0 < p < 1$. Then the number of trials up to and including the first success is a random variable X with probability mass function given by

$$p(x) = P(X = x) = (1-p)^{x-1}p, \quad x = 1, 2, 3, \ldots . \tag{4}$$

The random variable X is said to have a **geometric distribution with parameter p**. This is written $X \sim G(p)$.

The Ancient Greek mathematician Euclid knew about geometric progressions

Mathematically, the probabilities $P(X = 1)$, $P(X = 2)$, $P(X = 3)$, ... form a geometric progression: each probability in the sequence is a constant multiple (in this case $1 - p$) of the preceding one. This is the reason for the name 'geometric'.

Moreover, since $0 < p < 1$, it follows that $0 < 1 - p < 1$ also, so each consecutive probability is smaller than the preceding one. In other words, the geometric p.m.f. is always a decreasing function of x. This is

illustrated for two values of p in Figure 2. In Figure 2(a), the parameter p is equal to 0.8 – that is quite high: you would be unlikely to have to wait long for the first successful trial. In Figure 2(b), the value of the parameter p is much lower – the probability of success is only 0.3. In this case, you might have to wait quite a long time for the first success to occur. Whatever the value of p, however, the decreasing nature of the p.m.f. means that the most likely value of X is 1.

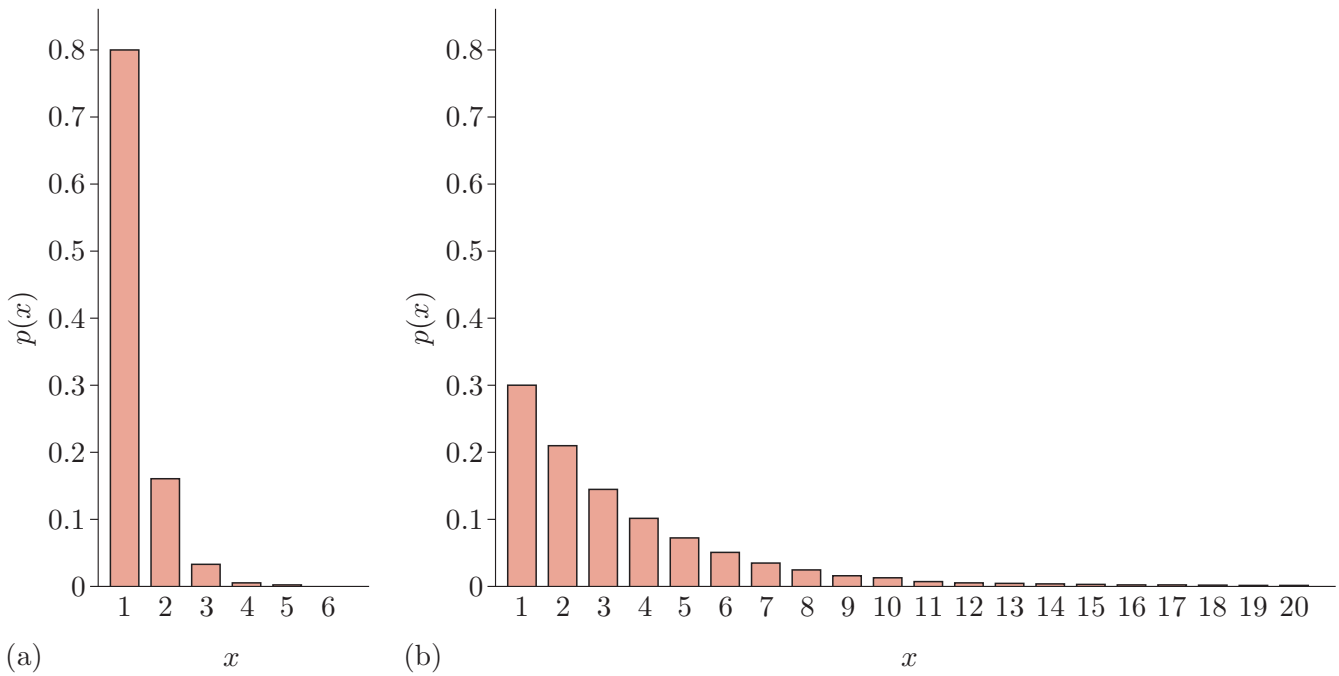


Figure 2 Two geometric p.m.f.s: (a) $p = 0.8$, (b) $p = 0.3$

The formula for the probability mass function of a geometric distribution is a simple one, so calculating probabilities is straightforward. Try the next two activities.

Activity 13 *Waiting to join in*

Suppose that, as in Example 10, you cannot start playing in a board game until you obtain a six on the roll of a six-sided die. Assume that the die is unbiased.

(a) What is the probability that you start playing with the first roll of the die?

(b) What is the probability that you start playing only with your second roll of the die?

(c) What is the probability that you start playing only with your eighth roll of the die?

Activity 14 *Defective silicon chips*

Suppose that the proportion of defective silicon chips leaving a factory's assembly line is 0.012. A quality inspector collects a random sample: she examines chips until she finds a defective one. What is the probability that she examines exactly ten chips?

The c.d.f. of a geometric distribution

In Subsection 2.2, it was noted that there is no convenient formula for the c.d.f. of a binomial random variable. However, it is possible to find a simple formula for the c.d.f. of a geometric random variable X. Moreover, the method for finding this simple formula is simple too if we use a little trick: first find an expression for the probability $P(X > x)$ and then obtain an expression for the c.d.f. $F(x) = P(X \leq x) = 1 - P(X > x)$.

Now, $P(X > x)$ is the probability that more than x trials are needed to obtain a success, or equivalently, that all the first x trials result in failure. This probability is equal to $(1-p)^x$. It follows that

$$F(x) = 1 - P(X > x) = 1 - (1-p)^x.$$

This is the formula we are seeking for the c.d.f. of a geometric distribution, stated in the box below.

The c.d.f. of a geometric distribution

If the random variable X has a geometric distribution with parameter p, then the c.d.f. of X is given by

$$F(x) = P(X \leq x) = 1 - (1-p)^x, \tag{5}$$

for $x = 1, 2, 3, \ldots$.

Example 12 *Still waiting to join in*

In Example 10 the situation was described in which you cannot start playing in a board game until you obtain a six on the roll of a six-sided die. From Activity 13, N, the number of rolls needed to obtain a six, assuming that the die is unbiased, has a geometric distribution: $N \sim G\left(\frac{1}{6}\right)$. So, for instance, the probability that you will have to roll the die at most three times may be obtained directly using the c.d.f. of N:

$$P(N \leq 3) = F(3) = 1 - (1-p)^3 = 1 - \left(\tfrac{5}{6}\right)^3 \simeq 0.421;$$

and the probability that you will have to roll the die more than three times is

$$P(N > 3) = 1 - P(N \leq 3) = 1 - F(3) = (1-p)^3 = \left(\tfrac{5}{6}\right)^3 \simeq 0.579.$$

Now use the c.d.f. to find the probabilities in the following activities.

Activity 15 *Still waiting to join in*

Suppose that, as in Example 12, you cannot start playing in a board game until you obtain a six on the roll of an unbiased six-sided die.

(a) What is the probability that you will have to roll the die at most four times?

(b) What is the probability that you will have to roll the die fewer than ten times?

(c) What is the probability that you will still be waiting to start playing after five rolls of the die?

Activity 16 *Defective silicon chips*

Suppose that, as in Activity 14, the proportion of defective silicon chips leaving a factory's assembly line is 0.012. The quality inspector collects a daily random sample. She examines sampled chips in the order in which they come off the assembly line until she finds one that is defective. She decides that she will halt production if fewer than six chips need examining. What percentage of her daily visits result in a halt in production?

3.2 Family patterns

The pattern of boys and girls in a family has received a lot of attention, both biological and statistical, over the last three hundred years. In 1710, the polymath John Arbuthnot, having examined parish records in London, noted that in each of the previous 82 years more boys than girls had been christened. He deduced that for 82 years more boys than girls had been born, and he proposed an explanation in terms of divine intervention: he argued that more male babies died than females, so 'To repair that loss, provident Nature, by the Disposal of its wise Creator, brings forth more Males than Females; and that in almost a constant proportion'.

John Arbuthnot (1667–1735)

The philosopher and probabilist Nicolaus Bernoulli (1687–1759) also noted the imbalance in the sexes, and commented that sex determination was like rolling a 35-sided die, with 18 faces marked 'boy' and 17 'girl'. Bernoulli's model assumes independence from child to child, and his estimate for the probability that a child born is a boy is $p = \frac{18}{35} \simeq 0.514$. If we let X be a random variable which takes the value 1 if a child is a boy and 0 for a girl, then

Nicolaus Bernoulli was a nephew of Jakob Bernoulli, whose name is attached to Bernoulli trials.

$$P(X = 0) = \tfrac{17}{35} \simeq 0.486, \qquad P(X = 1) = \tfrac{18}{35} \simeq 0.514.$$

So X, which represents the sex of a child, has a Bernoulli distribution with parameter $p = \frac{18}{35} \simeq 0.514$.

Activity 17 *Bernoulli's families*

Suppose that Nicolaus Bernoulli's model is a good one, and that the sexes of children born may be modelled as the outcomes of independent rolls of a 35-sided die with 18 faces marked 'boy' and the other 17 marked 'girl'.

(a) According to the model, what is the distribution of the number of boys in a family of n children?

(b) Find the probability that a family of four children:

(i) will all be girls;

(ii) will contain at least one boy;

(iii) will contain two boys and two girls.

One of the factors complicating the development of a satisfactory statistical model for family size and structure is that parents often impose their own 'stopping rules' for family limitation, depending on the number or distribution of boys and girls so far. For instance, among completed two-child families in a recent issue of *Who's Who in America* there was a striking excess of boy–girl and girl–boy sets, more than would be suggested by a simple binomial model: parents in that country (apparently) prefer to contrive their families to include at least one of each sex.

Activity 18 *Waiting for a girl*

Family limitation rules may be more extreme than that described above: for instance, a rule might be 'keep going until the first daughter is born, then stop'. Under this rule, possible family patterns for completed families (F for a daughter, M for a son) would be F, MF, MMF, MMMF, ... The number of children in a completed family of this type is a random variable.

(a) Assuming that Bernoulli's model holds, what is the probability distribution of this random variable?

(b) What proportion of completed families would contain the following?

(i) Exactly one child

(ii) Exactly two children

(iii) Exactly three children

(iv) Exactly four children

(c) What proportion of completed families would contain at least five children?

The next activity gives data on the birth order in families and comes from an investigation which was reported in 1963.

Activity 19 *Salt Lake City data*

Details were obtained on the sequence of the sexes of children in 116 458 families recorded in the archives of the Genealogical Society of Utah, part of The Church of Jesus Christ of Latter-day Saints in Salt Lake City. The records were examined to find the stage at which the first daughter was born in 7745 families where there was at least one daughter. The data are summarised in Table 3.

The Family History Library of The Church of Jesus Christ of Latter-day Saints in Salt Lake City

Table 3 First daughter

First daughter	Family structure	Family size	Frequency
Firstborn	F	1	3684
Secondborn	MF	2	1964
Thirdborn	MMF	3	1011
Fourthborn	MMMF	4	549
Later than fourth		≥ 5	537

(Source: James, W.H. (1987) 'The human sex ratio. Part I: A review of the literature', *Human Biology*, vol. 59, no. 5, pp. 721–52)

Find the sample relative frequencies for family size for the data in Table 3 and compare these values with the probabilities that you found in Activity 18, assuming Bernoulli's model is correct. Do the data appear to support a geometric model for family size?

Actually, a very considerable amount of data has been collected on the sex of children in families, and nearly all of these data suggest that the Bernoulli model is *not* a particularly good one. One theory suggests that the probability that a child is a boy probably varies from family to family, even if it averages 0.514 over a very large population. Some couples seem to have a preponderance of boys, and others of girls, and this occurs more frequently than could be explained simply by sampling variation under the binomial distribution. Even more tantalisingly, some statistical analyses of birth order seem to suggest that the independence assumption of Bernoulli's model is not valid: that is, that nature has a kind of memory, and the sex of a previous child affects the probability distribution of the sex of a subsequent child.

Exercises on Section 3

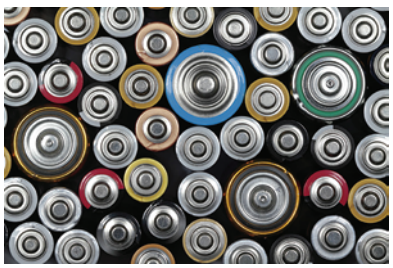
Exercise 5 *Calculating probabilities for geometric distributions*

(a) If $N \sim G(0.5)$, find the probability $P(N = 10)$.

(b) If $M \sim G\left(\frac{1}{3}\right)$, find the probability $P(M = 1)$.

(c) If $Q \sim G(0.1)$, find the probability $P(Q > 6)$.

(d) If $R \sim G(0.8)$, find the probability $P(R < 4)$.



Exercise 6 *Defective batteries*

The proportion of defective products in a battery factory is 0.02. A quality inspector tests batteries drawn independently at random from the assembly line.

(a) What is the probability that he will have to examine more than 20 batteries to obtain a faulty one?

(b) What is the probability that he will have to examine at least 50 batteries?

Exercise 7 *Waiting for a boy*

This exercise concerns the sizes of completed families for couples who stop having children when their first son is born. Suppose that Bernoulli's model is appropriate; in particular, the probability of a boy is $\frac{18}{35}$ and that of a girl is $\frac{17}{35}$.

(a) According to the model, what is the distribution of Y, the size of such a family?

(b) What proportion of completed families would contain the following?

(i) Two children

(ii) Four children

(c) What proportion of completed families would contain fewer than four children?

4 The Poisson probability distribution

In Unit 2, it was argued that discrete data often arise as *counts*, that is, as numbers of items or individuals with some attribute of interest. Three of the examples considered early in Unit 2 concern counts, but each has a slightly different nature which means that you already know about models for two of them, but not for the third. The first two examples concern binomial and geometric distributions, respectively; the third requires a different model. Let us give some further consideration to those three examples, with the first in a slightly different scenario.

Example 13 *Defective hinges*

A manufacturer produces hinges in batches of 1000, and is interested in the number of defective hinges, X say, in a batch. Then X is a count of the number of defective hinges, but it is one where there is a known upper limit to what that count might be: X takes any value from 0 (no defective hinges in a batch) up to 1000 (all hinges in a batch are defective). Because

the range of X is $\{0, 1, 2, \ldots, 1000\}$ and each hinge's defectiveness or otherwise is a Bernoulli trial which might be assumed to be independent of the defectiveness or otherwise of the other hinges in the batch, a model for the number of defective hinges is the binomial distribution. (Indeed, in such situations one might also say that interest centres more on the proportion of defective hinges than on their total number.)

Example 14 *Waiting to join in again*

Consider again the situation of a board game in which a player cannot join in until he or she has obtained a six on the roll of a die. The number of rolls necessary to obtain a six is a random variable N (say). It is a count and, unlike in Example 13, there is no upper limit to the value that the count can take (even though you would be extremely unlikely to have to wait, say, 1000 turns until you could start). Again, however, there is a Bernoulli trial structure underlying the count: N is the number of Bernoulli trials necessary in order to register a first success. As discussed in Section 3, an appropriate model for the distribution of N is the geometric distribution.

Example 15 *Yeast cells*

We started this unit by reconsidering the data on yeast cells that had been introduced in Unit 2. These data, given in Table 1 which is repeated here for convenience, are the number of yeast cells, X say, found in each of 400 very small squares on a microscope slide.

Table 4

Cells in a square, x	0	1	2	3	4	5
Frequency	213	128	37	18	3	1

Now, the data are once again counts, but there seems no natural sense in which the number of yeast cells in a square arises from some collection of Bernoulli trials. Combining this with the argument we made in the Introduction to this unit, that the range of X should not have a fixed finite upper bound, rules out the binomial distribution as a model for these data. The usual argument concerning waiting for a first success is also inapplicable, so that reason for using the geometric distribution is also not available. It seems that we need a new model for the probability distribution of count random variables with range $\{0, 1, 2, \ldots\}$.

In this section a probability distribution called the **Poisson distribution** is introduced as a model for data on the range $\{0, 1, 2, \ldots\}$. It is called the Poisson distribution after the French mathematician and physicist Siméon-Denis Poisson who, in 1837, introduced it in a scientific text on the subject of jury verdicts in criminal trials.

Siméon-Denis Poisson (1781–1840)

Now, it is possible to *derive* the Poisson distribution from considerations which relate to sequences of Bernoulli trials after all and, in fact, give the Poisson distribution a strong link to the binomial distribution. Thus, as was the case for the binomial and geometric distributions, justification for the use of the Poisson distribution is often based on consideration of the (probabilistic) 'mechanism' which describes how the data arose (from a collection of Bernoulli trials, as it happens, in all these cases).

A statistician's 'mechanism' is often not a physical mechanism

It turns out that the Poisson distribution emerges when we investigate what happens to the binomial, $B(n, p)$, distribution as the number of independent Bernoulli trials, n, becomes very large, and the probability of success, p, becomes very small. That is, we consider situations where, on the one hand, we are not very likely to obtain a success on any single trial – success is a rare event – but, on the other hand, we have many opportunities (trials) at which to obtain successes. Practical investigation of how this arises in the real world – and it does, in surprisingly many situations! – will be delayed until Unit 5. Mathematical investigation of how the Poisson distribution arises in this context, although optional, will be provided here, in order to avoid the mathematical formula for its probability mass function being entirely 'plucked out of thin air'.

λ is the Greek lower-case letter lambda, pronounced 'lamda'.

The key to the mathematical derivation of the Poisson distribution from the binomial is to write the binomial parameter p in the form $p = \lambda/n$ for some constant $\lambda > 0$. Remember that the sample size, n, is becoming large, so if $p = \lambda/n$, then p is becoming small at the same time. The derivation is included for interest and completeness in the following screencast; you will not be expected to reproduce any of the algebraic details. The result shown in the screencast is that as n becomes very large, the $B(n, \lambda/n)$ distribution tends to the Poisson(λ) distribution – the Poisson distribution with parameter λ – whose p.m.f. is provided in the box which follows shortly.

Screencast 3.2 Derivation of the Poisson distribution as a limit of binomial distributions (optional)

It is also valid to present the Poisson distribution directly by giving its p.m.f. with a view to using it as a model for data on $\{0, 1, 2, \ldots\}$ regardless of the mechanism underlying the data. This approach is often called *empirical* modelling. Of course, just because the Poisson distribution has the same range as the data doesn't mean that it is necessarily an appropriate model for the data. The Poisson distribution is a good model for the data only if it also reflects other features seen in the data: broadly speaking, this means if the Poisson p.m.f. quite closely resembles a relative frequency chart of the data.

Without further ado, the p.m.f. of the Poisson distribution is given in the following box.

The Poisson distribution

The random variable X is said to have a **Poisson distribution with parameter** λ if it has probability mass function

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots. \tag{6}$$

This is written $X \sim \text{Poisson}(\lambda)$. Here, $\lambda > 0$.

Notice that the Poisson distribution is indexed by a single parameter, λ, and that parameter can take any positive value. Interpretation of the role of λ will become clear in Units 4 and 5.

The Poisson p.m.f. involves the exponential function in the form of the factor $e^{-\lambda}$. It is a property of the exponential function that it is positive whatever the value of λ, and as $e^{-\lambda}$ is multiplied by other positive terms, $p(x) > 0$ for all $x = 0, 1, 2, \ldots$. Checking that $\sum_{x=0}^{\infty} p(x) = 1$ is done using the fact that

You will not be expected to reproduce the algebra in this paragraph.

$$e^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots.$$

This is the *Taylor series expansion* for e^{λ}. In some texts this expansion is used to define the exponential function.

Hence the sum of the probabilities is

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$$
$$= e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots\right) = e^{-\lambda}e^{\lambda} = 1,$$

as required.

Let's now see the Poisson distribution in action.

Example 16 *Insurance claims*

Suppose that the number of claims on a motor insurance policy over a 5-year period is a random variable X taking values $0, 1, 2, \ldots$. Its distribution can be modelled by a Poisson distribution with parameter $\lambda = 0.5$.

According to the Association of British Insurers, the UK insurance and long-term savings industry is the largest in Europe and the third largest in the world

Under the model, the probability that no claims are made on a motor insurance policy in 5 years is

$$P(X = 0) = p(0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-0.5} \simeq 0.607.$$

Also, the probability that exactly one claim is made is

$$P(X = 1) = p(1) = \frac{e^{-\lambda}\lambda^1}{1!} = e^{-0.5} \times 0.5 \simeq 0.303,$$

and the probability that more than one claim is made is

$$P(X > 1) = 1 - P(X \leq 1) = 1 - \{P(X = 0) + P(X = 1)\}$$
$$= 1 - e^{-0.5} - 0.5e^{-0.5} \simeq 0.090.$$

Activity 20 *Model for yeast cells*

A possible model for the distribution of the random variable X, the number of yeast cells found in a randomly chosen small square on a microscope slide, is Poisson(0.6825), that is, the Poisson distribution with $\lambda = 0.6825$.

(a) For a randomly chosen small square, what is the probability that there are no yeast cells present?

(b) For a randomly chosen small square, what is the probability that there are yeast cells present?

If X follows the Poisson(λ) distribution, then it is always the case that

$$P(X = 0) = p(0) = e^{-\lambda}.$$

In addition,

$$p(1) = \frac{e^{-\lambda}\lambda^1}{1!} = \lambda e^{-\lambda} = \lambda \times p(0),$$

$$p(2) = \frac{e^{-\lambda}\lambda^2}{2!} = \frac{\lambda}{2}\lambda e^{-\lambda} = \frac{\lambda}{2} \times p(1),$$

$$p(3) = \frac{e^{-\lambda}\lambda^3}{3!} = \frac{\lambda}{3}\frac{e^{-\lambda}\lambda^2}{2} = \frac{\lambda}{3} \times p(2),$$

and so on.

Activity 21 *Successive Poisson probabilities*

In the Poisson(λ) model, write $p(x)$ in terms of $p(x-1)$ for any $x = 1, 2, 3, \ldots$.

The relationships above allow you to calculate successive probabilities without repeated calculation of $e^{-\lambda}$: calculate $e^{-\lambda}$ once, to obtain $p(0)$, then multiply the result by λ to obtain $p(1)$, then multiply the result again by $\lambda/2$ to obtain $p(2)$, and so on.

Example 17 *More insurance claim probabilities*

Suppose again that the number of claims on a motor insurance policy over a 5-year period, X, follows the Poisson(0.5) distribution. In Example 16, we saw that

$$p(0) = e^{-0.5} \simeq 0.606\,530\,659 \simeq 0.607$$

and

$$p(1) = \lambda\, p(0) = \frac{1}{2}\, p(0) \simeq 0.303\,265\,329 \simeq 0.303.$$

Continuing in this way, we have

Keep full calculator accuracy as successive probabilities are computed.

$$p(2) = \frac{\lambda}{2}\,p(1) = \frac{1}{4}\,p(1) \simeq 0.075\,816\,332 \simeq 0.076,$$
$$p(3) = \frac{\lambda}{3}\,p(2) = \frac{1}{6}\,p(2) \simeq 0.012\,636\,055 \simeq 0.013,$$
$$p(4) = \frac{\lambda}{4}\,p(3) = \frac{1}{8}\,p(3) \simeq 0.001\,579\,506 \simeq 0.002,$$

and so on. In addition,

$$P(X \leq 4) = p(0) + p(1) + p(2) + p(3) + p(4) \simeq 0.9998.$$

If, on the other hand, we were interested only in $P(X \leq 4)$ (and not the individual probabilities), the main computational trick would be just to take the exponential term 'outside' and, again, calculate it only once:

$$\begin{aligned} P(X \leq 4) &= e^{-0.5} + e^{-0.5}\frac{0.5}{1} + e^{-0.5}\frac{(0.5)^2}{2!} + e^{-0.5}\frac{(0.5)^3}{3!} + e^{-0.5}\frac{(0.5)^4}{4!} \\ &= e^{-0.5}\left(1 + 0.5 + \frac{0.25}{2} + \frac{0.125}{6} + \frac{0.0625}{24}\right) \\ &\simeq e^{-0.5} \times 1.64844 \simeq 0.9998. \end{aligned}$$

Activity 22 *Yeast cell probabilities*

Consider again the distribution of X, the number of yeast cells found in a randomly chosen small square on a microscope slide.

(a) For the Poisson(0.6825) distribution:

(i) Calculate $p(0)$, $p(1)$, $p(2)$, $p(3)$ and $p(4)$.

(ii) What is $P(X > 4)$?

(b) The observed data on the number of yeast cells per square were given in Table 1 and repeated in Example 15. Find the sample relative frequencies for these data and compare these values with the probabilities that you found in part (a) under the Poisson(0.6825) distribution. Do the data appear to support this Poisson model for numbers of yeast cells?

Using the relationship between successive Poisson probabilities helps a little in alleviating the calculational burden of evaluating the c.d.f. of a Poisson distribution by adding together individual probabilities. As for the binomial distribution – but unlike the geometric distribution – the c.d.f. of a Poisson random variable does not have a convenient mathematical form. But again, the computer can help. In the same way as you used Minitab to calculate probabilities for binomial distributions in conjunction with Subsection 2.2 (Chapter 7 of Computer Book A), so Minitab can also be used to calculate probabilities for Poisson distributions. You will have the opportunity to use this facility when you reach Unit 5 of the module.

The Poisson probability mass function exhibits two main shapes, depending on the value of λ, as illustrated in Figure 3. When $\lambda < 1$, the Poisson p.m.f., like the geometric p.m.f., is decreasing (and zero is the most likely value of X); see Figure 3(a) for an example. On the other hand, if $\lambda > 1$, the Poisson p.m.f. has an up-then-down shape; see Figure 3(b) for an example of this kind.

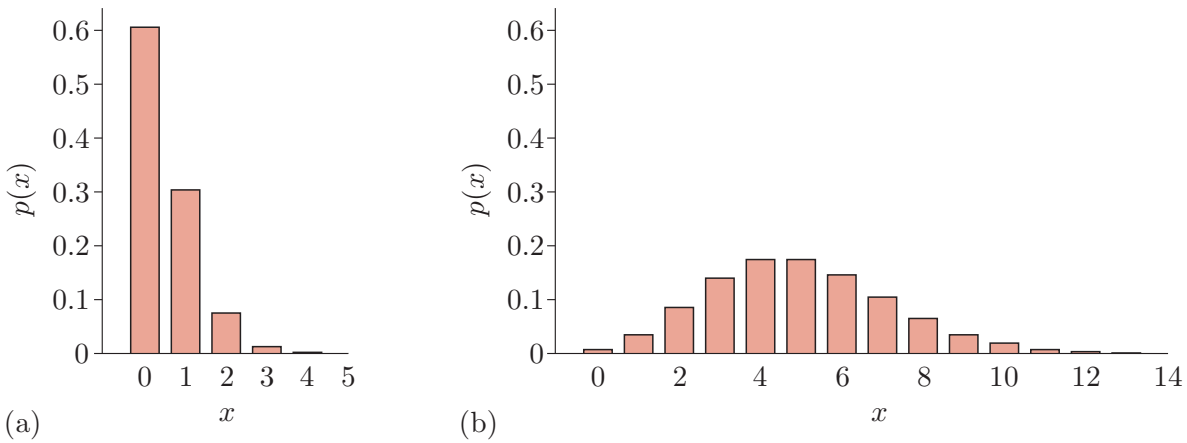


Figure 3 Two Poisson p.m.f.s: (a) $\lambda = 0.5$, (b) $\lambda = 5$

It is, in fact, the result of Activity 21 that for the Poisson(λ) distribution,

$$p(x) = \frac{\lambda}{x} \times p(x-1), \quad x = 1, 2, 3, \ldots,$$

that tells us about the shape of a Poisson p.m.f. The appropriate connections – including the case of $\lambda = 1$ – are made in Screencast 3.3.

Screencast 3.3 The shapes of Poisson distributions

Exercises on Section 4

Exercise 8 *Calculating probabilities for Poisson distributions*

(a) If $N \sim \text{Poisson}(5)$, find the probability $P(N = 6)$.

(b) If $M \sim \text{Poisson}(0.7)$, find the probability $P(M = 1)$.

(c) If $Q \sim \text{Poisson}(2)$, find the probability $P(Q > 1)$.

(d) If $R \sim \text{Poisson}\left(\frac{1}{3}\right)$, find the probability $P(R \leq 2)$.

Exercise 9 *Hurricanes*

In the Northern Atlantic Ocean, hurricanes occur from early June to late November, with a peak in late August and September. Suppose that the number of Atlantic hurricanes each year can be modelled by a Poisson distribution with $\lambda = 6$. Assuming the Poisson model to be a good one, find the following probabilities.

(a) What is the probability of there being no Atlantic hurricanes in a year?

(b) What is the probability of there being exactly six Atlantic hurricanes in a year?

(c) What is the probability of there being more than four Atlantic hurricanes in a year?

5 Two models for uniformity

In this section, we consider uniform distributions, which are models for situations where all the possible outcomes of an experiment are equally likely to occur. In Subsection 5.1, a discrete probability model of this type is introduced. Despite the discrete nature of all the other models considered in this unit, it proves convenient to also introduce a continuous version of this model in this unit; it is discussed in Subsection 5.2. A short subsection on the role of uniform distributions in statistics, including an important special case of the continuous uniform distribution, completes the section and the unit in Subsection 5.3.

5.1 The discrete uniform distribution

We begin by revisiting two examples from Unit 2.

Example 18 *The score on an unbiased die*

When an unbiased six-sided die is rolled, it comes to rest displaying any one of its six faces with equal probability. At any given roll of the die, the score obtained on the uppermost face is a random variable Y. As you saw in Unit 2, the probability mass function for Y is

$$p(y) = 1/6, \quad y = 1, 2, \ldots, 6.$$

Example 19 *Roulette wheels*

European roulette wheels have 37 equal-sized compartments numbered $0, 1, \ldots, 36$. If a wheel is fair, then Z, the number of the compartment in which the ball comes to rest, is equally likely to be any in this range. Its probability mass function is therefore

$$p(z) = 1/37, \quad z = 0, 1, \ldots, 36.$$

In Unit 2, you considered certain specific probabilities, not the whole distribution of Z.

These two examples are of random variables having a **discrete uniform distribution**. The list of possible values that each random variable X can take (the range of X) is given as a set of integers with stated lower and upper limits (1 and 6 in Example 18, 0 and 36 in Example 19); *and no possible value within those limits is more probable than any other possible value.*

You should be comfortable with changes of notation like this: in Example 18 the random variable was called Y, and in Example 19 it was called Z.

A general definition of the discrete uniform distribution is as follows.

The discrete uniform distribution

Suppose that the range of the random variable X is $m, m+1, \ldots, n-1, n$. Then X is said to have a **discrete uniform distribution** with parameters m and n if it has probability mass function

$$p(x) = \frac{1}{n-m+1}, \quad x = m, m+1, \ldots, n. \tag{7}$$

A distribution of uniforms

Again, there is a whole family of discrete uniform distributions: the indexing parameters in this case are m (the minimum attainable value) and n (the maximum attainable value). Often the value of m is 1, in which case

$$p(x) = \frac{1}{n}, \quad x = 1, 2, \ldots, n.$$

For example, this is the case for the score on an unbiased die, when $m = 1$ and $n = 6$, giving

$$p(x) = \frac{1}{6}, \quad x = 1, 2, \ldots, 6,$$

as in Example 18. In Example 19, however, $m = 0$ and $n = 36$, since the 37 compartments on a European roulette wheel are labelled $0, 1, \ldots, 36$. For general n, if $m = 0$, the p.m.f. of the discrete uniform distribution is

$$p(x) = \frac{1}{n+1}, \quad x = 0, 1, \ldots, n.$$

The p.m.f.s associated with Examples 18 and 19, illustrated in Figure 4, show the typical shapes of all discrete uniform distributions: they are all 'flat' in the sense that $p(x)$ is constant for all x in its range. (Figure 4(a) is the same as Figure 3 of Unit 2.)

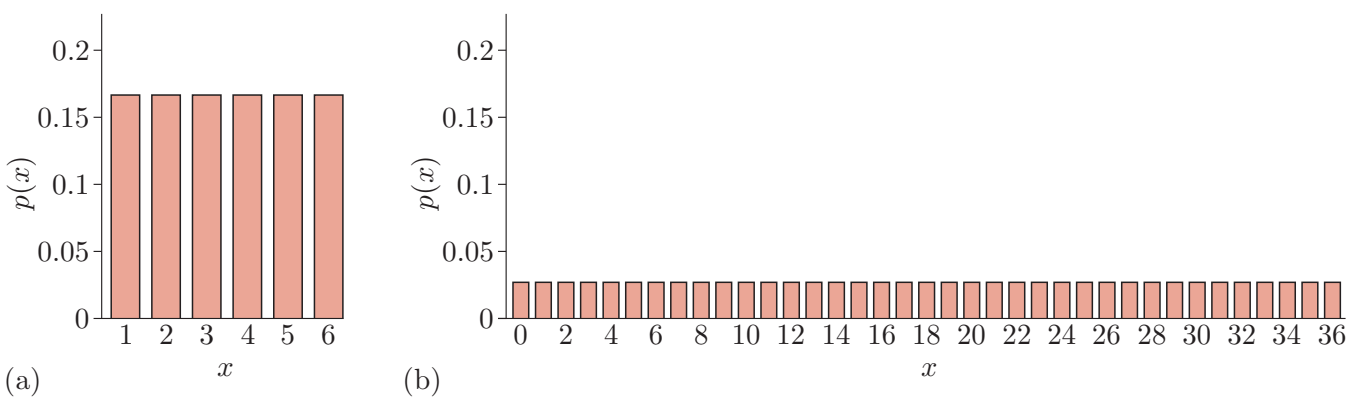


Figure 4 Two discrete uniform p.m.f.s: (a) $m = 1$, $n = 6$, (b) $m = 0$, $n = 36$

The cumulative distribution function of a discrete uniform distribution has quite a straightforward formula.

The c.d.f. of a discrete uniform distribution

The c.d.f. of a discrete uniform distribution with parameters m and n is

$$F(x) = P(X \leq x) = \frac{x - m + 1}{n - m + 1}, \quad x = m, m+1, \ldots, n. \tag{8}$$

Activity 23 *The c.d.f.s of discrete uniform distributions*

(a) Prove that the c.d.f. of a discrete uniform distribution with parameters m and n is

$$F(x) = \frac{x - m + 1}{n - m + 1}, \quad x = m, m+1, \ldots, n.$$

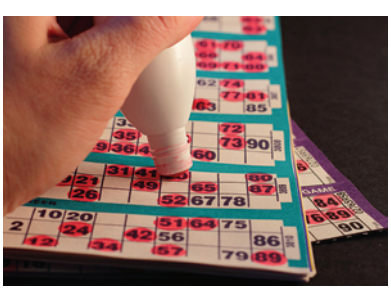
(b) What is the c.d.f. for the following?

(i) $m = 1$

(ii) $m = 0$

Activity 24 *Bingo numbers*

A common form of the gambling game Bingo uses 90 balls that are numbered $1, 2, \ldots, 90$. Balls are drawn at random, and players win by being the first to mark off particular sets of numbers on cards. Let Y be the number of the first ball that is drawn.



(a) Write down the p.d.f. and c.d.f. of Y.

(b) What are $P(Y \leq 20)$ and $P(Y \leq 35)$? Hence determine $P(21 \leq Y \leq 35)$.

So far in this subsection, we have been able to assert the appropriateness of the discrete uniform distribution as a model in several situations purely by consideration of the nature of those situations. There are other cases, however, where we have data on a finite range of integer values and only a more questionable idea that the probabilities of each value might be (at least approximately) the same. A rather quirky example of this kind is considered next.

Example 20 *Month of death of royal descendants*

The data in Table 5 (overleaf) give the month of death (January = 1, February = 2, ..., December = 12) for 82 descendants of Queen Victoria; they all died of natural causes. A bar chart of the data is given in Figure 5 (overleaf).

Queen Victoria (centre, holding baby) and many members of her family

Table 5 Month of death of royal descendants

Month	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	13	4	7	10	8	4	5	3	4	9	7	8

(Source: Andrews, D. and Herzberg, A. (1985) *Data*, New York, Springer, p. 429)

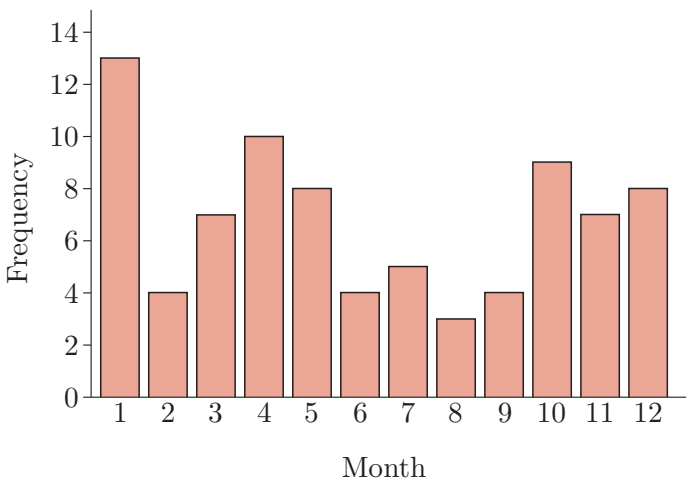


Figure 5 A bar chart of the deaths in each month

So is uniformity a reasonable model here? Well, perhaps not: at a first glance, the data seem to suggest that the summer months (months 6, 7, 8 and 9) are less likely to include a death than the winter months, apart from February (month 2). This could be explained by deaths being weather-related and, at least partly, by February being a short month. On the other hand, perhaps this kind of pattern can be attributed purely to random variation. Perhaps a uniform distribution is reasonable, and if, as here, we have just 82 observations from such a distribution, a sample bar chart might quite reasonably look like the one we have.

You should now explore the questions raised in Example 20 by turning to your computer.

Refer to Chapter 8 of Computer Book A to explore the uniformity or otherwise of the royal deaths data.

As a postscript to your investigation in Chapter 8 of Computer Book A, Figure 5 does not reflect the fact that the months of January and December are adjacent. This does not matter when we are just considering whether or not the distribution is uniform, as we have been doing, but it would be very important to take into account if we were to decide that the uniform distribution is not a good model, and wish to explore other models for these data.

5.2 The continuous uniform distribution

In this last but one subsection of the unit, we switch attention from discrete distributions to a particular distribution for continuous data. The reason is that the continuous distribution of interest has much in common with the discrete distribution just investigated; both are distributions for outcomes which, in some appropriate sense, are all 'equally likely'. The following examples describe some situations for which such a continuous statistical model might be relevant.

Example 21 *Faulty cable*

Faults in underground cables cause degradation, or even complete loss, of the signal carried by the cable. When this happens, the cable needs to be repaired. In the absence of any indication of where the fault might be, the repair company has to search the cable until the fault is located – this is just as likely to be near the end, near the beginning or in the middle of the cable. The distance searched to locate the fault is a random variable, and a factor in the cost of the repair.

Example 22 *Fractional ages*

When someone dies, we usually say something like 'he was 92' or 'she was 64'; in everyday life, knowing a person's age to the nearest year is sufficient for many purposes. Actuaries working in life insurance, however, need more precise ages at death. As part of their modelling of ages at death, actuaries often split ages into whole years plus the 'fractional age', the *proportion* of the year from the deceased's final birthday to the time of death. As part of actuarial models, the fractional age is often considered to be equally likely to be any value between 0 and 1.

The central idea in these examples is that of 'no preferred value'. The probability model for which there is no preferred value for continuous data between two bounds is known as the **continuous uniform distribution**. If there is no preferred value in the range of a continuous random variable X, then the height of the p.d.f. of X on a graph must be constant over the range of X.

Activity 25 *The p.d.f. of a continuous uniform distribution*

Suppose that X can take values between a and b ($a < b$); then a sketch of its p.d.f. will look like that in Figure 6 (overleaf).

This p.d.f. is of the form $f(x) = h$, for $a < x < b$. Use the fact that the total area under the graph of a probability density function is 1 to find the value of the constant h in terms of a and b. Hence write down the p.d.f. of X.

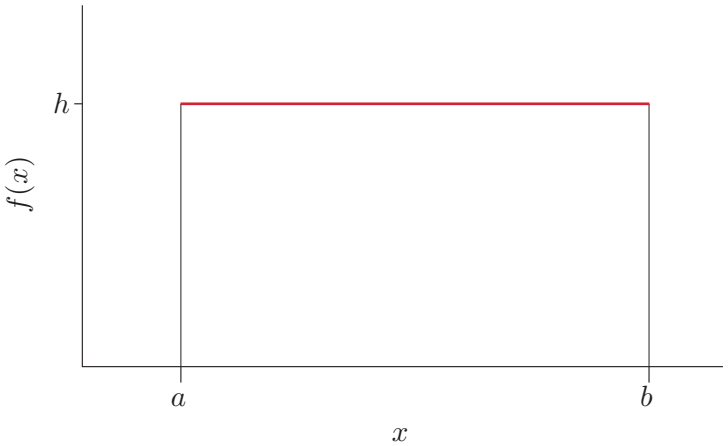


Figure 6 No preferred value between a and b

The definition of a continuous uniform distribution and the formula for its p.d.f. are given in the box below.

The continuous uniform distribution

The continuous random variable X which is equally likely to take any value between two stated bounds a and b $(a < b)$ is said to be **uniformly distributed over the interval** (a, b). The probability density function of X is

$$f(x) = \frac{1}{b-a}, \quad a < x < b. \tag{9}$$

This is written $X \sim U(a, b)$.

The continuous uniform probability density function was shown in Figure 6 with $h = 1/(b-a)$. The distribution is sometimes called the rectangular distribution, for the obvious reason. Also, recall from Unit 2 that the probability that a continuous random variable takes any specific value is effectively zero. We have chosen, as is most standard, to leave the endpoints a and b out of the range of X above. However, we could have included either or both of a and b in the range and it would have made no difference. This is because the probability of getting exactly a or getting exactly b is effectively zero.

We therefore use the open interval notation (a, b) rather than the closed interval notation $[a, b]$ for the range of the distribution.

Activity 26 *Model for faulty cable*

Suppose there is one fault in a cable that is 40 metres long and that, as in Example 21, the fault is equally likely to be anywhere along the cable. Let X denote the distance in metres along the cable to the fault. What model should we use for the distribution of X? What is the p.d.f. of X in this case?

Using integration, we can use the p.d.f. of X to determine the probability that X takes a value in a specified interval. From Subsection 3.2 of Unit 2, for any continuous random variable X,

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)\,dx.$$

Remember that, for any continuous distribution, $P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2)$ and that you can likewise disregard the difference between '$<$' and '$\leq$' in any similar calculation.

Example 23 *Faulty cable: probability of the fault's location*

Suppose again that there is one fault in a cable that is 40 metres long. The cable is particularly difficult to access between 20 and 30 metres from its start. What is the probability that the fault is in the section of the cable that is particularly difficult to access?

With X denoting the distance in metres along the cable to the fault, and assuming that $X \sim U(0, 40)$ (as in Activity 26), we need to calculate $P(20 \leq X \leq 30)$. This is given by

$$\int_{20}^{30} f(x)\,dx = \int_{20}^{30} \frac{1}{40}\,dx = \left[\frac{x}{40}\right]_{20}^{30} = \frac{30}{40} - \frac{20}{40} = \frac{1}{4}.$$

Activity 27 *Another faulty cable*

Let X be the distance (in metres) along a 100-metre cable to the one fault that the cable contains, where $X \sim U(0, 100)$. Use integration to find the probability that the fault is between 50 m along the cable and the far end of the cable.

The probabilities found in Example 23 and Activity 27 agree with common sense:

- In Example 23 the length of cable from 20 m to 30 m is 10 m. This is one-quarter of the 40 m cable, so the probability that it contains the fault is one-quarter.
- In Activity 27 the length of cable from 50 m to 100 m is 50 m. This is one-half of the 100 m cable, so the probability that it contains the fault is one-half.

Indeed, probabilities associated with the continuous uniform distribution agree with common sense quite generally, as you can check in the following activity.

Activity 28 *Common sense probabilities*

Let $X \sim U(a, b)$ and evaluate the probability that X lies in the interval (c, d) where $a \leq c < d \leq b$. Comment on the interpretation of the result that you obtain.

Integration of its p.d.f. also yields the cumulative distribution function of a continuous uniform distribution.

Activity 29 *The c.d.f. of a continuous uniform distribution*

Use the definition of the c.d.f. of a continuous random variable, namely,

See Subsection 4.3 of Unit 2.

$$F(x) = P(X \leq x) = \int_a^x f(y)\,dy \quad \text{for } a < x < b,$$

to obtain the formula for $F(x)$ when $X \sim U(a, b)$.

The result of Activity 29 is worth stressing.

The c.d.f. of a continuous uniform distribution

If X has a continuous uniform distribution, $X \sim U(a, b)$, then its c.d.f. is

$$F(x) = P(X \leq x) = \frac{x-a}{b-a}, \quad a < x < b. \tag{10}$$

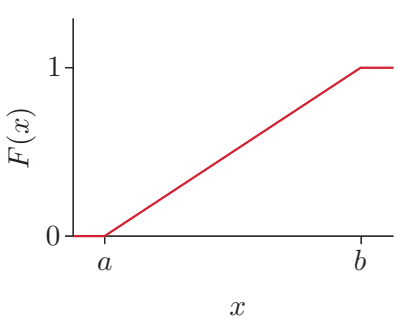


Figure 7 The continuous uniform c.d.f.

Notice that this c.d.f. has a particularly simple form: it is linear in x. That is, for the $U(a, b)$ distribution, $F(x)$ is a straight line joining the value $F(a) = (a-a)/(b-a) = 0$ to the value $F(b) = (b-a)/(b-a) = 1$. See Figure 7.

Recall also from Subsection 4.3 of Unit 2 that for any continuous random variable X, probabilities of lying within intervals can be written in terms of the c.d.f. as

$$P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2) = F(x_2) - F(x_1).$$

When $X \sim U(a, b)$ and $a \leq c < d \leq b$, this allows us to confirm the result of Activity 28:

$$P(c < X < d) = F(d) - F(c) = \frac{d-a}{b-a} - \frac{c-a}{b-a}$$
$$= \frac{d-a-c+a}{b-a} = \frac{d-c}{b-a}.$$

Activity 30 *Faulty cable: probabilities from the c.d.f.*

Suppose again that X represents the position, in metres, of the single fault in a cable of length 100 m, and assume that $X \sim U(0, 100)$.

(a) What is the c.d.f. of X?

(b) Use the c.d.f. of X to obtain the following probabilities.

(i) The probability that X is less than 25 m

(ii) The probability that X is more than 75 m

(iii) The probability that X is between 15 m and 35 m

Activity 31 *Waiting to see the dentist*

Observation over many months has shown that the time a patient spends in the waiting room at a dentist's surgery varies between 2 minutes and 20 minutes; no time in this range seems more or less likely than any other time. The random variable W represents the time a patient has to wait.

(a) What distribution may be used to model the random variable W?

(b) Write down the c.d.f. of W.

(c) What is the probability that on her next visit a patient will have to wait for the following times?

(i) Less than five minutes

(ii) More than a quarter of an hour

5.3 How useful is the uniform distribution?

The main uses of most probability distributions, like the binomial, the geometric and the Poisson, are directly in the modelling of random variables of interest based on samples of data, as seen in earlier sections of this unit. This is less so for either uniform distribution. But they are none the less extremely important distributions in statistics. This is for two further reasons.

The first is that uniform distributions can be used to express a person's *beliefs* in an approach to statistics called Bayesian statistics, which is not covered in this module. Suffice it to say that when those beliefs are very

vague and 'everything seems equally likely' (before data are collected), the uniform distributions can have an important role (as what are called prior distributions).

The second important role for the uniform distributions is in computer *simulation*, the body of techniques that you have already been using in your computer work, in which the computer is used to mimic, and hence more efficiently explore, the workings and outcomes of real experiments. In particular, in Chapter 6 of Computer Book A we discussed using computer 'random number generators' that generate sequences of 'pseudo-random' integers. These numbers are indistinguishable from sequences of truly random numbers, by which we actually mean observations from the discrete uniform distribution. However, the range of this *discrete* uniform distribution is truly enormous, so much so that, by dividing by an integer little larger than the largest integer obtained, we actually think of random numbers as observations from the *continuous* uniform distribution on the range $(0, 1)$. That is, a pseudo-random number, although strictly speaking of rational form, is indistinguishable from a realisation of the continuous random variable V where $V \sim U(0, 1)$.

The continuous uniform distribution on $(0, 1)$ is sometimes called the *standard uniform distribution* (or possibly more often, just $U(0, 1)$, pronounced 'U 0 1'). Using results from Subsection 5.2, you can obtain the basic properties of the standard uniform distribution yourself, in the next activity.

Activity 32 *The standard uniform distribution*

The random variable V has the standard uniform distribution, $V \sim U(0, 1)$.

(a) Write down the p.d.f. of V.

(b) Write down the c.d.f. of V.

(c) What is the probability that V lies between 0.1 and 0.8?

On account of the importance of the standard uniform distribution, its basic properties are summarised in the following box.

The standard uniform distribution

For the continuous random variable V which has the standard uniform distribution, $V \sim U(0, 1)$:

- the p.d.f. of V is
$$f(v) = 1, \quad 0 < v < 1,$$
- the c.d.f. of V is
$$F(v) = v, \quad 0 < v < 1.$$

The standard uniform distribution is, in fact, the basis of all simulation. This bold claim can be made because of the remarkable fact that random variables following other probability distributions can all be simulated – sometimes straightforwardly, sometimes using sophisticated algorithms – using just standard uniform random variables. It is beyond the scope of this module to go into this in general here, but we will briefly explore a link with one other distribution: the discrete uniform distribution with range $\{1, 2, \ldots, k\}$, particularly having in mind cases where k is quite small.

In your computer work on Units 2 and 3 you have already used random variables generated from other distributions derived from standard uniform random variables.

Activity 33 *Simulating the score on an unbiased die*

Suppose you have available an observation, v, of the random variable V which has the standard uniform distribution. You wish to simulate the outcome of a single roll of an unbiased six-sided die. The p.m.f. of the latter random variable, Y say, is

$$p(y) = 1/6, \quad y = 1, 2, \ldots, 6.$$

How do you think you might obtain a value for Y with the correct distribution, given the value of v?

This was done by the computer in Activity 24 of Chapter 5 of Computer Book A.

The intuitive method for simulating the outcome of a single roll of an unbiased six-sided die works in general, as shown in the following.

Suppose again that, as in Activity 33, we have available an observation, v, of the random variable V which has the standard uniform distribution. And now we wish to simulate a value for the random variable, X, which has the discrete uniform distribution with p.m.f.

$$p(x) = 1/k, \quad x = 1, 2, \ldots, k.$$

Similarly to Activity 33, an intuitive method for doing so would be to divide $(0, 1)$ into k equal intervals, $(0, 1/k), [1/k, 2/k), \ldots, [(k-1)/k, 1)$, and to associate with each of these intervals the values $1, 2, \ldots, k$, respectively, for X. That is, given $V = v$, set

$$X = \begin{cases} 1 & \text{if } 0 < v < 1/k \\ 2 & \text{if } 1/k \leq v < 2/k \\ \vdots & \qquad\vdots \\ k & \text{if } (k-1)/k \leq v < 1. \end{cases}$$

In the unlikely event that computer rounding of values of v results in exactly 0 or exactly 1, simply throw these values away and try again.

Activity 34 *Proving that it works*

Prove that the method just described in the text works by showing that $P(X = x) = 1/k$ for any x in $\{1, 2, \ldots, k\}$.

Exercises on Section 5

Exercise 10 *The score on a dodecahedral die*

A regular dodecahedral die has 12 faces labelled $1, 2, \ldots, 12$. The random variable X represents the score on the face on which the die lands when it is rolled.

(a) Write down the probability mass function of X.

(b) Write down the cumulative distribution function of X.

(c) Find $P(X < 8)$ using the c.d.f. of X.

Exercise 11 *Watch batteries*

Replacement watch batteries are claimed to last 'between one and two years'.

(a) Assuming that any length of time between one and two years is equally likely, suggest a model for T, the length of time, in years, replacement watch batteries last.

(b) Write down the c.d.f. of T for the model you suggested in part (a).

(c) According to your model, what is the probability that a replacement watch battery will last for 15 months or less?

(d) According to your model, what is the probability that a replacement watch battery will last between 15 and 21 months?

Summary

In this unit, you have been introduced to a number of specific probability distributions as models for data. Five of these are models for discrete data:

- the Bernoulli distribution to describe the outcome of a single Bernoulli trial
- the binomial distribution to describe the total number of successes in a set of independent Bernoulli trials each with the same probability of success
- the geometric distribution to describe the number of independent Bernoulli trials up to and including the first success
- the Poisson distribution as a model for count data
- the discrete uniform distribution for use when each of a range of integer values is equally likely to occur.

In addition, you have met the continuous analogue of the discrete uniform distribution in which any value within an interval of values is equally likely to occur.

Learning outcomes

After you have worked through this unit, you should be able to:

- understand the meaning of the term Bernoulli trial, which describes a single statistical experiment for which there are two possible outcomes, often referred to as 'success' and 'failure'
- appreciate that if the outcome of one trial does not influence the outcome of another, then the trials are said to be independent
- appreciate that many probability distributions may be regarded as members of a family of distributions, and that a family of distributions is indexed by one or more parameters
- identify the distribution (Bernoulli, binomial or geometric) of a random variable associated with a sequence of independent Bernoulli trials each with the same probability of success, and specify the parameter(s) of the distribution
- predict the shapes of the probability functions of the distributions considered in this unit, given values for their parameters
- calculate probabilities for binomial and Poisson random variables using their p.m.f.s
- calculate probabilities for geometric and uniform random variables using either their p.m.f.s or their c.d.f.s
- calculate binomial probabilities using Minitab
- appreciate the degree of random variation associated with some of these probability distributions.

Solutions to activities

Solution to Activity 1

First substitute $x = 1$ in the equation. We obtain

$$p(1) = p^1(1-p)^{1-1} = p(1-p)^0 = p.$$

Then writing $x = 0$ gives

$$p(0) = p^0(1-p)^{1-0} = 1-p.$$

Solution to Activity 2

(a) The probability that a woman, randomly selected from the population of 6503 women, has passed the menopause is

$$\frac{591}{6503} \simeq 0.091.$$

(b) Let the random variable X take the value 1 if a woman in this population has passed the menopause and 0 otherwise. Taking p to be exactly 0.091, $X \sim$ Bernoulli(0.091), so

$$p(x) = \begin{cases} 0.909 & x = 0 \\ 0.091 & x = 1 \end{cases}$$

or

$$p(x) = (0.091)^x\,(0.909)^{1-x}, \quad x = 0, 1.$$

(Remember that it is important to specify the range of the random variable.)

Solution to Activity 3

(a) The total number of people of Indian descent should not be modelled by a binomial distribution because the independence assumption does not hold: if one person in a UK household is of Indian descent, the probability that the other person chosen from the same household is of Indian descent is increased.

(b) The total number of obese children in the sample could be modelled by a binomial distribution. Independence of different children across a number of inner-city estates and the same probability of obesity for any randomly chosen child would appear to be reasonable assumptions.

(c) The total number of young people who play football should not be modelled by a binomial distribution because p is not constant: for instance, the probability of playing football is going to be different depending on the gender of the young people.

Solution to Activity 4

(a) The number of successes in 15 trials can be any integer value between 0 (if there are no successes) and 15 (if every trial is a success). Hence the range of the random variable X is $\{0, 1, 2, \ldots, 15\}$.

(b) The number of successes in n trials can be any integer value between 0 (if there are no successes) and n (if every trial is a success). Hence the range of the random variable X is $\{0, 1, 2, \ldots, n\}$.

Solution to Activity 5

(a) $\displaystyle\binom{5}{3} = \frac{5!}{3!\,2!} = \frac{1\times 2\times 3\times 4\times 5}{(1\times 2\times 3)(1\times 2)} = \frac{4\times 5}{2} = 10.$

(b) $\displaystyle\binom{7}{1} = \frac{7!}{1!\,6!} = \frac{1\times 2\times 3\times 4\times 5\times 6\times 7}{(1)(1\times 2\times 3\times 4\times 5\times 6)} = 7.$

(c) $\displaystyle\binom{8}{0} = \frac{8!}{0!\,8!} = \frac{8!}{1\times 8!} = 1.$

(d) $\displaystyle\binom{6}{4} = \frac{6!}{4!\,2!} = \frac{1\times 2\times 3\times 4\times 5\times 6}{(1\times 2\times 3\times 4)(1\times 2)} = \frac{5\times 6}{2} = 15.$

Many calculators can find values of factorials and binomial coefficients directly. If your calculator has either of these facilities, then make sure you know how to use them. You may find these facilities useful when calculating binomial probabilities.

Solution to Activity 6

(a) $\displaystyle P(X=4) = \binom{6}{4}(0.4)^4(1-0.4)^{6-4} = \frac{6!}{4!\,2!}(0.4)^4(0.6)^2$

$$= 15(0.4)^4(0.6)^2 = 0.13824 \simeq 0.138.$$

(b) $\displaystyle P(V=2) = \binom{8}{2}(0.3)^2(1-0.3)^{8-2} = \frac{8!}{2!\,6!}(0.3)^2(0.7)^6$

$$= 28(0.3)^2(0.7)^6 \simeq 0.296.$$

Solution to Activity 7

(a) The number dropping out in the placebo group, Y, is binomial: $Y \sim B(6, 0.14)$.

(i) The probability that all six drop out is

$$P(Y=6) = \binom{6}{6}(0.14)^6(1-0.14)^0 = \frac{6!}{6!\,0!}(0.14)^6$$
$$= (0.14)^6 \simeq 0.0000075.$$

(ii) The probability that none of the six drops out is

$$P(Y=0) = \binom{6}{0}(0.14)^0(1-0.14)^6 = \frac{6!}{0!\,6!}(0.86)^6$$
$$= (0.86)^6 \simeq 0.405.$$

(iii) The probability that exactly two drop out is

$$P(Y=2) = \binom{6}{2}(0.14)^2(1-0.14)^4 = \frac{6!}{2!\,4!}(0.14)^2(0.86)^4$$
$$= 15(0.14)^2(0.86)^4 \simeq 0.161.$$

(b) The assumption of independence, in this case, is essentially saying that whether a patient drops out of the placebo group is unaffected by whether or not other patients in the group drop out. Sometimes patients are unaware of others' progress in this sort of trial; but if that is not the case, then it is possible that a large drop in numbers would discourage others from continuing in the study. Similarly, even in the absence of obvious beneficial effects, patients might offer mutual encouragement to persevere. In such circumstances, the independence assumption breaks down.

Solution to Activity 8

Two of the three distributions are skew, but the first (for which $p = 1/2$) is symmetric. The p.m.f. in Figure 1(b) (for which $p > 1/2$) is left-skew; the p.m.f. in Figure 1(c) (for which $p < 1/2$) is right-skew.

Solution to Activity 9

(a) Since the Bernoulli distribution is that of the outcome of a single Bernoulli trial with success probability p, and the binomial distribution is that of the total number of successes in n independent Bernoulli trials with success probability p, the Bernoulli(p) distribution is the $B(1, p)$ distribution (the binomial distribution with $n = 1$).

(b) When $n = 1$, the binomial p.m.f. from Equation (3) is

$$p(x) = \binom{1}{x}p^x(1-p)^{1-x} = \frac{1!}{x!\,(1-x)!}\,p^x(1-x)^{1-x}, \quad x = 0, 1.$$

This reduces to Equation (1), which is $p(x) = p^x(1-p)^{1-x}$, $x = 0, 1$, because $x!$ and $(1-x)!$ take only the values 0! and 1!, both of which are 1.

Solution to Activity 10

(a) $P(W \leq 1) = P(W = 0) + P(W = 1)$

$$= \binom{7}{0}(0.6)^0(0.4)^{7-0} + \binom{7}{1}(0.6)^1(0.4)^{7-1}$$
$$= 1(0.4)^7 + 7(0.6)(0.4)^6 \simeq 0.0016 + 0.0172 \simeq 0.019.$$

(b) $P(Y=0) + P(Y=3) = \binom{3}{0}\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^{3-0} + \binom{3}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^{3-3}$

$$= 1\left(\frac{2}{3}\right)^3 + 1\left(\frac{1}{3}\right)^3 = \frac{8}{27} + \frac{1}{27} = \frac{1}{3}.$$

(c) $P(X > 4) = P(X = 5) + P(X = 6)$

$$= \binom{6}{5}(0.8)^5(0.2)^{6-5} + \binom{6}{6}(0.8)^6(0.2)^{6-6}$$

$$= 6(0.8)^5(0.2) + 1(0.8)^6 \simeq 0.3932 + 0.2621 \simeq 0.655.$$

Solution to Activity 11

$$\sum_{x=0}^{n} p(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x}.$$

But by the binomial theorem with $a = p$, $b = 1 - p$,

$$\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + 1 - p)^n = 1^n = 1,$$

as required.

Solution to Activity 12

(a) The random variable X takes the value 1 if the first trial results in a success: $P(X = 1) = p$.

(b) Success occurs for the first time at the second trial if the first trial is a failure and the second is a success: $P(X = 2) = (1 - p) \times p$.

(c) Success occurs for the first time at the third trial if the first two trials are both failures and the third is a success:
$P(X = 3) = (1 - p) \times (1 - p) \times p = (1 - p)^2 p$.

(d) A pattern is emerging here. The random variable X takes the value x if the first $(x - 1)$ trials are failures and these are followed by a success:

$$P(X = x) = (1 - p)^{x-1} p, \quad x = 1, 2, 3, \ldots.$$

Solution to Activity 13

The number of rolls, N, needed to start playing, has a geometric distribution with parameter $p = \frac{1}{6}$.

(a) $P(N = 1) = p = \frac{1}{6} \simeq 0.167$.

(b) $P(N = 2) = (1 - p)\, p = \frac{5}{6} \times \frac{1}{6} = \frac{5}{36} \simeq 0.139$.

(c) $P(N = 8) = (1 - p)^7 p = \left(\frac{5}{6}\right)^7 \times \frac{1}{6} \simeq 0.047$.

Solution to Activity 14

If identification of a defective chip is regarded as a success, then X, the number of chips the inspector examines, has a geometric distribution with parameter $p = 0.012$. So

$$P(X = 10) = (0.988)^9 \times 0.012 \simeq 0.011.$$

Solution to Activity 15

Let N be the number of rolls of the die needed to obtain a six; then $N \sim G\left(\frac{1}{6}\right)$.

(a) The probability required is

$$P(N \leq 4) = F(4) = 1 - (1-p)^4 = 1 - \left(\tfrac{5}{6}\right)^4 \simeq 0.518.$$

(b) Here we require

$$P(N < 10) = P(N \leq 9) = F(9) = 1 - (1-p)^9 = 1 - \left(\tfrac{5}{6}\right)^9 \simeq 0.806.$$

(c) The probability required here is

$$P(N > 5) = 1 - P(N \leq 5) = 1 - F(5) = (1-p)^5 = \left(\tfrac{5}{6}\right)^5 \simeq 0.402.$$

Solution to Activity 16

The number of chips the inspector examines, X, has a geometric distribution with parameter $p = 0.012$. The probability required is

$$P(X < 6) = P(X \leq 5) = F(5) = 1 - (1-p)^5 = 1 - (0.988)^5 \simeq 0.059.$$

So approximately 6% of her daily visits involve a halt in production.

Solution to Activity 17

(a) If Y is the number of boys in a family of n children, then Y follows a binomial distribution, $Y \sim B\left(n, \frac{18}{35}\right)$.

(b) If Y is the number of boys in a family of four children, then $Y \sim B\left(4, \frac{18}{35}\right)$. We use Equation (3) to answer each of the parts of the question.

(i) The probability that all four children are girls is

$$P(Y = 0) = \binom{4}{0} \left(\tfrac{18}{35}\right)^0 \left(\tfrac{17}{35}\right)^{4-0} = \left(\tfrac{17}{35}\right)^4 \simeq 0.056.$$

(ii) The probability that at least one child is a boy is

$$P(Y \geq 1) = 1 - P(Y = 0) \simeq 0.944.$$

(iii) The probability that two children are boys and two are girls is

$$P(Y = 2) = \binom{4}{2} \left(\tfrac{18}{35}\right)^2 \left(\tfrac{17}{35}\right)^{4-2} = 6 \left(\tfrac{18}{35}\right)^2 \left(\tfrac{17}{35}\right)^2 \simeq 0.374.$$

Solution to Activity 18

(a) If N is the number of children in a completed family, then $N \sim G\left(\frac{17}{35}\right)$. Notice that here a 'success' is having a girl, which has probability $p = \frac{17}{35}$.

(b) Using Equation (4), the proportions of completed families with each of one, two, three and four children are found as follows.

(i) $P(N = 1) = p = \frac{17}{35} \simeq 0.486.$

(ii) $P(N = 2) = (1 - p)p = \frac{18}{35} \times \frac{17}{35} \simeq 0.250.$

(iii) $P(N = 3) = (1 - p)^2 p = (\frac{18}{35})^2 \times \frac{17}{35} \simeq 0.128.$

(iv) $P(N = 4) = (1 - p)^3 p = (\frac{18}{35})^3 \times \frac{17}{35} \simeq 0.066.$

(c) Using Equation (5), the proportion of completed families with at least five children is given by

$$P(N \geq 5) = 1 - P(N \leq 4) = 1 - F(4)$$
$$= (1 - p)^4 = \left(\tfrac{18}{35}\right)^4 \simeq 0.070.$$

Solution to Activity 19

The sample relative frequencies are obtained by dividing the sample frequencies by 7745. This gives 0.476, 0.254, 0.131, 0.071 and 0.069, respectively, correct to three decimal places. These do not differ greatly from the proportions calculated in Activity 18, assuming Bernoulli's model, which were 0.486, 0.250, 0.128, 0.066 and 0.070, respectively,

You were not asked to do this, but it is also interesting to note that successive frequencies in the data are in the ratios

$$\frac{1964}{3684} \simeq 0.533, \quad \frac{1011}{1964} \simeq 0.515, \quad \frac{549}{1011} \simeq 0.543,$$

which are all just over a half. The frequencies themselves form a geometric progression (approximately). So it looks as though a geometric model might be reasonable. However, more formal methods of assessing the quality of fit of a model are required in order to decide whether or not a model is a good one; one such method is discussed in a later unit of this module.

Solution to Activity 20

(a) We require $P(X = 0)$. This is

$$P(X = 0) = p(0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-0.6825} \simeq 0.505.$$

(b) We require $P(X \geq 1)$. This is

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-0.6825} \simeq 0.495.$$

Solution to Activity 21

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{\lambda}{x} \times \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \frac{\lambda}{x} \times p(x-1).$$

(If you watched Screencast 3.2, you should recognise that this is the relationship between successive probabilities that led to the declaration that the limiting distribution was Poisson.)

Solution to Activity 22

(a) (i) Retaining full calculator accuracy in the underlying calculations, we have (similar to Example 17)

$$p(0) = e^{-0.6825} \simeq 0.505,$$

and then

$$p(1) = \lambda\, p(0) = 0.6825\, p(0) \simeq 0.345,$$
$$p(2) = \frac{\lambda}{2}\, p(1) = \frac{0.6825}{2}\, p(1) \simeq 0.118,$$
$$p(3) = \frac{\lambda}{3}\, p(2) = \frac{0.6825}{3}\, p(2) \simeq 0.027$$

and

$$p(4) = \frac{\lambda}{4}\, p(3) = \frac{0.6825}{4}\, p(3) \simeq 0.005.$$

(ii)
$$\begin{aligned} P(X > 4) &= 1 - P(X \leq 4) \\ &= 1 - \{p(0) + p(1) + p(2) + p(3) + p(4)\} \\ &= 1 - e^{-0.6825}\left(1 + \frac{0.6825}{1!} + \frac{0.6825^2}{2!} + \frac{0.6825^3}{3!} + \frac{0.6825^4}{4!}\right) \\ &\simeq 0.001. \end{aligned}$$

(b) The sample relative frequencies are obtained by dividing the sample frequencies by 400. They are given in the following table along with the probabilities obtained above, all values being given correct to three decimal places (and as there was one square containing 5 yeast cells, the sample proportion of values of X greater than 4 was 0.003).

Table 6

Cells in a square, x	0	1	2	3	4	5
Frequency	213	128	37	18	3	1
Relative frequency	0.533	0.320	0.093	0.045	0.008	0.003
Poisson probability	0.505	0.345	0.118	0.027	0.005	0.001

These sample relative frequencies are broadly similar to the probabilities obtained above under the Poisson(0.6825) model. It seems that perhaps the Poisson model is not an unreasonable model for these data.

Solution to Activity 23

(a) For $x = m, m+1, \ldots, n$,

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= P(X = m) + P(X = m+1) + P(X = m+2) + \cdots + P(X = x) \\ &= \underbrace{\frac{1}{n-m+1} + \frac{1}{n-m+1} + \frac{1}{n-m+1} + \cdots + \frac{1}{n-m+1}}_{x-m+1 \text{ terms}} \\ &= \frac{x-m+1}{n-m+1}. \end{aligned}$$

(b) (i) When $m = 1$,

$$F(x) = \frac{x}{n}, \quad x = 1, 2, \ldots, n.$$

(ii) When $m = 0$,

$$F(x) = \frac{x+1}{n+1}, \quad x = 0, 1, \ldots, n.$$

Solution to Activity 24

(a) As $n - m + 1 = 90 - 1 + 1 = 90$, the p.d.f. of Y is

$$p(y) = \frac{1}{90}, \quad y = 1, 2, \ldots, 90.$$

As $y - m + 1 = y - 1 + 1 = y$, the c.d.f. of Y is

$$F(y) = \frac{y}{90}, \quad y = 1, 2, \ldots, 90.$$

(b) $$P(Y \leq 20) = F(20) = \frac{20}{90} = \frac{2}{9} \simeq 0.222$$

and

$$P(Y \leq 35) = F(35) = \frac{35}{90} = \frac{7}{18} \simeq 0.389.$$

Hence

$$P(21 \leq Y \leq 35) = F(35) - F(20) = \frac{35 - 20}{90} = \frac{15}{90} = \frac{1}{6} \simeq 0.167.$$

Solution to Activity 25

The total area under the graph in Figure 6 is $(b - a) \times h$. For this area to be 1, we need $(b - a)h = 1$ and hence $h = 1/(b - a)$. So the p.d.f. of X is

$$f(x) = \frac{1}{b - a}, \quad a < x < b.$$

Solution to Activity 26

Because the fault is equally likely to be anywhere between $a = 0$ and $b = 40$, $X \sim U(0, 40)$ and the p.d.f. of X is

$$f(x) = \frac{1}{40}, \quad 0 < x < 40.$$

Solution to Activity 27

The probability that the fault is between 50 m and 100 m along the cable is

$$\int_{50}^{100} f(x)\,dx = \int_{50}^{100} \frac{1}{100}\,dx = \left[\frac{x}{100}\right]_{50}^{100} = \frac{100}{100} - \frac{50}{100} = \frac{1}{2}.$$

Solution to Activity 28

$$P(c < X < d) = \int_c^d \frac{1}{b-a}\,dx = \left[\frac{x}{b-a}\right]_c^d = \frac{d}{b-a} - \frac{c}{b-a} = \frac{d-c}{b-a}.$$

This makes sense because the probability that X lies in the interval (c, d) is the length of the interval (c, d), that is, $d - c$, expressed as a proportion of the length of the entire range (a, b), which is $b - a$.

Solution to Activity 29

$$F(x) = \int_a^x \frac{1}{b-a}\,dy = \left[\frac{y}{b-a}\right]_a^x = \frac{x}{b-a} - \frac{a}{b-a} = \frac{x-a}{b-a}.$$

Solution to Activity 30

(a) With $a = 0$ and $b = 100$, the c.d.f. is

$$F(x) = \frac{x-0}{100-0} = \frac{x}{100}, \quad 0 < x < 100.$$

(b) (i) $P(X < 25) = F(25) = \dfrac{25}{100} = \dfrac{1}{4}.$

(ii) $P(X > 75) = 1 - P(X \leq 75) = 1 - F(75) = 1 - \dfrac{75}{100} = \dfrac{1}{4}.$

(iii) $P(15 < X < 35) = F(35) - F(15) = \dfrac{35}{100} - \dfrac{15}{100} = \dfrac{1}{5}.$

Solution to Activity 31

(a) The random variable W may be modelled by a continuous uniform distribution over the interval $2 < w < 20$, that is, $W \sim U(2, 20)$.

(b) The c.d.f. of W is

$$F(w) = \frac{w-2}{20-2} = \frac{w-2}{18}, \quad 2 < w < 20.$$

(c) (i) The probability that a patient will have to wait for less than five minutes is

$$P(W < 5) = F(5) = \frac{3}{18} = \frac{1}{6} \simeq 0.167.$$

(ii) The probability that a patient will have to wait for more than a quarter of an hour is

$$\begin{aligned} P(W > 15) &= 1 - P(W \leq 15) = 1 - F(15) \\ &= 1 - \frac{13}{18} = \frac{5}{18} \simeq 0.278. \end{aligned}$$

Solution to Activity 32

(a) Using Equation (9), the p.d.f. of V is

$$f(v) = \frac{1}{1-0} = 1, \quad 0 < v < 1.$$

(b) Using Equation (10), the c.d.f. of V is

$$F(v) = \frac{v-0}{1-0} = v, \quad 0 < v < 1.$$

(c) $P(0.1 < V < 0.8) = F(0.8) - F(0.1) = 0.8 - 0.1 = 0.7.$

Solution to Activity 33

The random variable Y has six possible outcomes, each with equal probability, $1/6$, of occurring. It would seem reasonable to divide the range $(0, 1)$ into six equal parts, $(0, 1/6), [1/6, 1/3), \ldots, [5/6, 1)$, and to associate with each of these intervals a value for Y. So, for example, we might associate the interval $(0, 1/6)$ with the value 1, the interval $[1/6, 1/3)$ with the value 2, and so forth. Then whichever of the intervals the observed value of v lies in defines a value for Y; for example, if $1/3 \leq v < 1/2$, then Y can be set equal to 3.

Solution to Activity 34

The key to the proof is that the event $X = x$ is the same as the event $(x-1)/k \leq V < x/k$. Let F denote the c.d.f. of $V \sim U(0, 1)$. It follows that

$$\begin{aligned} P(X = x) &= P\left(\frac{x-1}{k} \leq V < \frac{x}{k}\right) = F\left(\frac{x}{k}\right) - F\left(\frac{x-1}{k}\right) \\ &= \frac{x}{k} - \frac{x-1}{k} = \frac{x - x + 1}{k} = \frac{1}{k}, \end{aligned}$$

as required, for any x in $\{1, 2, \ldots, k\}$.

Solutions to exercises

Solution to Exercise 1

$X \sim$ Bernoulli(0.04). Its p.m.f. can also be written

$$p(x) = (0.04)^x\,(0.96)^{1-x}, \quad x = 0, 1.$$

Solution to Exercise 2

(a) $\binom{6}{1} = \dfrac{6!}{1!\,5!} = \dfrac{1\times 2\times 3\times 4\times 5\times 6}{(1)(1\times 2\times 3\times 4\times 5)} = \dfrac{6}{1} = 6.$

(b) $\binom{9}{2} = \dfrac{9!}{2!\,7!} = \dfrac{1\times 2\times 3\times 4\times 5\times 6\times 7\times 8\times 9}{(1\times 2)(1\times 2\times 3\times 4\times 5\times 6\times 7)} = \dfrac{8\times 9}{1\times 2} = 36.$

(c) $\binom{7}{4} = \dfrac{7!}{4!\,3!} = \dfrac{1\times 2\times 3\times 4\times 5\times 6\times 7}{(1\times 2\times 3\times 4)(1\times 2\times 3)} = \dfrac{5\times 6\times 7}{1\times 2\times 3} = 35.$

Solution to Exercise 3

(a) $P(X = 3) = \binom{4}{3}(0.6)^3(0.4)^{4-3} = 4(0.6)^3(0.4) = 0.3456.$

(b) $P(Y = 2) = \binom{7}{2}(0.2)^2(0.8)^{7-2} = 21(0.2)^2(0.8)^5 \simeq 0.275.$

(c)
$$\begin{aligned} P(Z > 6) &= P(Z = 7) + P(Z = 8) \\ &= \binom{8}{7}(0.75)^7\,(0.25)^{8-7} + \binom{8}{8}(0.75)^8\,(0.25)^{8-8} \\ &= 8\,(0.75)^7\,(0.25) + (0.75)^8 \simeq 0.367. \end{aligned}$$

Solution to Exercise 4

(a) If X is the number of defective items in the sample, then $X \sim B(20, 0.05)$, so

$$\begin{aligned} P(X \le 1) &= P(X = 0) + P(X = 1) \\ &= \binom{20}{0}(0.05)^0(0.95)^{20-0} + \binom{20}{1}(0.05)^1(0.95)^{20-1} \\ &= (0.95)^{20} + 20(0.05)(0.95)^{19} \simeq 0.736. \end{aligned}$$

(b) If Y is the number of shots that hit the centre of the target in ten shots, then $Y \sim B(10, 0.9)$, so

$$P(Y = 8) = \binom{10}{8}(0.9)^8(0.1)^{10-8} = 45(0.9)^8(0.1)^2 \simeq 0.194.$$

(c) If Z is the number of matches the tennis player wins out of five matches, then $Z \sim B(5, 0.7)$, so

$$\begin{aligned} P(Z \geq 3) &= P(Z = 3) + P(Z = 4) + P(Z = 5) \\ &= \binom{5}{3} (0.7)^3 (0.3)^{5-3} + \binom{5}{4} (0.7)^4 (0.3)^{5-4} \\ &\quad + \binom{5}{5} (0.7)^5 (0.3)^{5-5} \\ &= 10(0.7)^3(0.3)^2 + 5(0.7)^4(0.3) + (0.7)^5 \simeq 0.837. \end{aligned}$$

Solution to Exercise 5

(a) $P(N = 10) = (0.5)^9 \times 0.5 = (0.5)^{10} \simeq 0.001$.

(b) $P(M = 1) = \frac{1}{3}$.

(c) $P(Q > 6) = 1 - P(Q \leq 6) = 1 - F(6) = (0.9)^6 \simeq 0.531$.

(d) $P(R < 4) = P(R \leq 3) = F(3) = 1 - (0.2)^3 = 0.992$.

Solution to Exercise 6

In this case X, the number of batteries tested, has a geometric distribution with parameter $p = 0.02$.

(a) The probability that the inspector has to examine more than 20 batteries is

$$P(X > 20) = 1 - P(X \leq 20) = 1 - F(20) = (0.98)^{20} \simeq 0.668.$$

(b) The probability that the inspector has to examine at least 50 batteries is

$$P(X \geq 50) = 1 - P(X \leq 49) = 1 - F(49) = (0.98)^{49} \simeq 0.372.$$

Solution to Exercise 7

(a) $Y \sim G\left(\frac{18}{35}\right)$.

(b) The proportions of completed families with each of two and four children are given as follows.

(i) $P(Y = 2) = \frac{17}{35} \times \frac{18}{35} \simeq 0.250$.

(ii) $P(Y = 4) = \left(\frac{17}{35}\right)^3 \times \frac{18}{35} \simeq 0.059$.

(c) The proportion of completed families with fewer than four children is given by

$$P(Y < 4) = P(Y \leq 3) = F(3) = 1 - \left(\frac{17}{35}\right)^3 \simeq 0.885.$$

Solution to Exercise 8

(a) $P(N = 6) = \dfrac{e^{-5}5^6}{6!} \simeq 0.146.$

(b) $P(M = 1) = \dfrac{e^{-0.7}0.7}{1} \simeq 0.348.$

(c) $P(Q > 1) = 1 - P(Q \leq 1) = 1 - \{p(0) + p(1)\}$
$= 1 - e^{-2}\left(1 + \dfrac{2}{1}\right) \simeq 0.594.$

(d) $P(R \leq 2) = p(0) + p(1) + p(2) = e^{-1/3}\left(1 + \dfrac{1/3}{1!} + \dfrac{(1/3)^2}{2!}\right) \simeq 0.995.$

Solution to Exercise 9

Let X denote the number of Atlantic hurricanes in a year.

(a) $P(X = 0) = e^{-6} \simeq 0.002.$

(b) $P(X = 6) = \dfrac{e^{-6}6^6}{6!} \simeq 0.161.$

(c) $P(X > 4) = 1 - P(X \leq 4) = 1 - \{p(0) + p(1) + p(2) + p(3) + p(4)\}$
$= 1 - e^{-6}\left(1 + \dfrac{6}{1!} + \dfrac{6^2}{2!} + \dfrac{6^3}{3!} + \dfrac{6^4}{4!}\right)$
$= 1 - e^{-6}\left(1 + 6 + \dfrac{36}{2} + \dfrac{216}{6} + \dfrac{1296}{24}\right)$
$= 1 - e^{-6} \times 115 \simeq 0.715.$

Solution to Exercise 10

(a) By Equation (7), the p.m.f. of X is

$$p(x) = 1/12, \quad x = 1, 2, \ldots, 12.$$

(b) By Equation (8), the c.d.f. of X is

$$F(x) = x/12, \quad x = 1, 2, \ldots, 12.$$

(c) $P(X < 8) = P(X \leq 7) = F(7) = 7/12.$

Solution to Exercise 11

(a) A model for the time T that a replacement watch battery will last is a continuous uniform distribution over the interval $1 < t < 2$; that is, $T \sim U(1, 2)$.

(b) The c.d.f. of T is (using Equation (10))

$$F(t) = \frac{t - 1}{2 - 1} = t - 1, \quad 1 < t < 2.$$

(c) Converting 15 months to 1.25 years, the required probability is

$$P(T \leq 1.25) = F(1.25) = 1.25 - 1 = 0.25.$$

(d) Converting 21 months to 1.75 years, the required probability is

$$P(1.25 < T < 1.75) = F(1.75) - F(1.25)$$
$$= (1.75 - 1) - (1.25 - 1) = 0.5.$$

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 143: © NASA

Page 146: Taken from: http://kienthuc.net.vn/me-be/ly-thu-qua-trinh-tao-mau-o-thai-nhi-341713.html

Page 148: © dolgachov / www.123rf.com

Page 149: © Cathy Yeulet / www.123rf.com

Page 150: © Randy Glasbergen

Page 151: © Alan Gignoux / www.Dreamstime.com

Page 153 top: Taken from: https://lastinieblasdelamente.wordpress.com/category/noticias-curiosas/

Page 153 bottom: © Stuart Miles / www.123rf.com

Page 156: © Djomas / www.shutterstock.com

Page 161: © 06photo / www.shutterstock.com

Page 162: Hzenilc / https://commons.wikimedia.org/wiki/File:Euclid%27s_Elements_1573_Edition.JPG This file is licensed under the Creative Commons Attribution-ShareAlike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 166: © Sergey Novikov / www.123rf.com

Page 168 top: © Martin Konopka / www.123rf.com

Page 168 bottom: Phil Holmes / www.123rf.com

Page 170: © Andrey Armyagov / www.123rf.com

Page 172: © John Gomez / www.istock.photo.com

Page 176: © gstockstudio / www.123rf.com

Page 177: © Melanie Braun / www.123rf.com

Page 179: Richard Ash / www.flickr.com This file is licensed under the Creative Commons Attribution-ShareAlike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 183: © bowdenimages / www.istockphoto.com

Page 186: © Blackregis / www.istockphoto.com

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.